Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

MATEUS SILVA PEDROSO

Modelos Aditivos Generalizados para Locação, Escala e Forma na Contagem de Insetos Vetores de Huanglongbing (HLB)

MARINGÁ – PARANÁ
2022

MATEUS SILVA PEDROSO

Modelos Aditivos Generalizados para Locação, Escala e Forma na Contagem de Insetos Vetores de Huanglongbing (HLB)

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do centro de ciências exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientador: Prof.ª Dr.ª Terezinha Aparecida Guedes
Coorientador: Prof. Dr. Willian Luís de Oliveira

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

MARINGÁ – PARANÁ

2022

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Pedroso, Mateus Silva

P372m Modelos aditivos generalizados para locação, escala e forma na contagem de Insetos Vetores de Huanglongbing (HLB) / Mateus Silva Pedroso. -- Maringá, PR, 2022.
89 f.: il. color., figs., tabs.

Orientadora: Profa. Dra. Terezinha Aparecida Guedes.
Coorientador: Prof. Dr. Willian Luís de Oliveira.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2022.

1. Bioestatística - Modelos Aditivos Generalizados. 2. Bioestatística - Modelos lineares generalizados Mistos. 3. Modelos de Regressão. 4. Dados Longitudinais. 5. Medidas Repetidas. I. Guedes, Terezinha Aparecida, orient. II. Oliveira, Willian Luís de, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 23.ed. 519.5

Jane Lessa Monção - CRB 9/1173

MATEUS SILVA PEDROSO

Modelos Aditivos Generalizados para Locacao, Escala e Forma na Contagem de Insetos Vetores de Huanglongbing (HLB)

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Profª. Dra. Terezinha Aparecida Guedes
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. William Mário de Carvalho Nunes
Universidade Estadual de Maringá – PGA/UEM

Prof. Dr. Vanderly Janeiro
Universidade Estadual de Maringá – PBE/UEM

Maringá, 13 de abril de 2022.

Agradecimentos

Quero agradecer a todos, familiares e amigos, que estiveram do meu lado e me apoiaram nos momentos difíceis, me incentivando a continuar nessa desafiadora jornada.

Agradecimentos especiais a minha orientadora Prof.ª Terezinha, dona de um coração gigante e que sempre demonstrou empatia, respeito e paciência, fundamental para que dia após dia esse trabalho fosse construído. Ao professor Willian, meu coorientador, que sempre esteve a disposição para me dar o suporte necessário.

Estendo meus agradecimentos também a todos que, de alguma forma, contribuíram no processo. A todos os professores do programa pelos quais passei e tenho profunda gratidão e respeito. Agradeço à Universidade Estadual de Maringá pela oportunidade e à CAPES por subsidiar financeiramente a pesquisa.

Deus abençoe a todos!

Mateus Silva Pedroso

“Nós nos transformamos naquilo que praticamos com frequência. A perfeição, portanto, não é um ato isolado. É um hábito.”

(Aristóteles)

Resumo

O estudo do controle e manejo eficiente contra pragas e doenças tem fundamental importância na atividade agrícola, em especial, à citricultura brasileira, uma vez que o país é líder na produção mundial de laranja doce e suco concentrado de laranja. O greening dos citros (Huanglongbing - HLB) é uma das doenças mais devastadoras na produção de citros e tem sido responsável por grandes prejuízos aos produtores e, desta forma, seu combate se faz necessário. Para isso, no mapeamento da doença, pode-se recorrer aos métodos estatísticos para modelar a infestação de insetos vetores, que são responsáveis pela transmissão da doença. Neste contexto, tem-se geralmente a geração de estudos longitudinais com medidas repetidas, no qual é avaliado o comportamento do indivíduo (uma planta infectada, por exemplo) ao longo do tempo. Além disso, é necessário que se tenha o suporte estatístico para lidar com as respostas correlacionadas que surgem de estudos desse tipo, e também, para modelar respostas não normais como contagens e proporções. Na literatura, encontra-se a aplicação de modelos lineares generalizados mistos (MLGM) e modelos aditivos para locação, escala e forma (GAMLSS) como principais metodologias. Nesse sentido, o objetivo do trabalho foi modelar a presença de insetos vetores em pomares comerciais de laranja com incidência de HLB associado a possíveis fatores explicativos, utilizando as metodologias MLGM e GAMLSS e propondo uma discussão entre tais abordagens. A variável resposta considerada foi a contagem de insetos capturados no experimento e, portanto, foram selecionadas as distribuições de probabilidade Poisson Generalizada, Poisson Normal Inversa, Binomiais Negativas Tipo I e Tipo II, Família Binomial Negativa, Beta Binomial Negativa, Sichel e Delaporte. Os resultados indicaram que a distribuição Binomial Negativa, pertencente a família exponencial de distribuições, teve o melhor desempenho dentre as distribuições testadas. Com relação às metodologias, constatou-se que o modelo linear generalizado misto não foi capaz de modelar adequadamente devido a heterogeneidade da variância. Assim, utilizou-se a flexibilidade da metodologia GAMLSS para modelar o parâmetro de escala da distribuição e com isso, obteve-se um modelo adequado e com os pressupostos satisfeitos.

Palavras-chave: Modelos de regressão; Dados longitudinais; Medidas repetidas; Efeitos aleatórios.

Abstract

The study of efficient control and management against pests and diseases is of fundamental importance in agricultural activity, especially in Brazilian citrus industry, since the country is the leader world's production of sweet orange and concentrated orange juice. Citrus greening (Huanglongbing - HLB) is one of the most devastating diseases in citrus production and has been responsible for great losses to producers and, therefore, its combat is necessary. For this, in mapping the disease, statistical methods can be used to model the infestation of vector insects, which are responsible for the transmission of the disease. In this context, there is generally the generation of longitudinal studies with repeated measures, in which the behavior of the individual (an infected plant, for example) is evaluated over time. Furthermore, it is necessary to have statistical support to deal with correlated responses that arise from studies of this type, and also to model non-normal responses such as counts and proportions. In the literature, it is found the application of generalized linear mixed models (GLMM) and generalized additive models for location, scale and shape (GAMLSS) as main methodologies. In this sense, the objective of this work was to model the presence of vector insects in commercial orange orchards with an increase in HLB associated with possible explanatory factors, using the MLGM and GAMLSS methodologies and proposing a discussion between such approaches. The response variable considered was the number of insects captured in the experiment and, therefore, the Generalized Poisson, Inverse Normal Poisson, Type I and Type II Negative Binomials, Negative Binomial Family, Negative Binomial Beta, Sichel and Delaporte were selected. The results indicated that the Negative Binomial distribution, belonging to the exponential family of distributions, had the best performance among the distributions tested. Regarding the methodologies, it was found that the mixed generalized linear model was not able to model properly due to the heterogeneity of the variance. Thus, the flexibility of the GAMLSS methodology was used to model the scale parameter of the distribution and with that, an adequate model was obtained, with the assumptions satisfied.

Keywords: Regression Models; Longitudinal data; Repeated measures; Random effects.

Lista de ilustrações

Figura 2.1 – Localização das áreas experimentais. 20
Figura 2.2 – Coleta das armadilhas adesivas amarelas fixadas em pomar comercial de laranja, em que: 1, 2 e 3 – Armadilha em campo; 4 – Armadilha sequenciada para ser identificada; 5 – Armadilhas protegidas e identificadas; 6 – Separação das armadilhas por talhão e pontos; 7 – Análise em lupa binocular e anotação de contagem de insetos vetores; 8 – Armadilha vista com lupa; 9 – Identificação de *Diaphorina citri*. 21
Figura 2.3 – Exemplificação da transformação de uma variável observada, y, no resíduo quantílico aleatorizado, r (Stasinopoulos et al., 2017, p. 421). 37
Figura 2.4 – Exemplo de *worm plot* de um modelo bem ajustado (Stasinopoulos et al., 2017, p. 427). 39
Figura 2.5 – Exemplo de *worm plot* pela função `rqres.plot()` para distribuições de probabilidade discretas. 40
Figura 2.6 – Exemplo da função `plot()` de um modelo GAMLSS (Stasinopoulos et al., 2017, p. 424). 40
Figura 3.1 – Gráficos de barras da distribuição de contagem de Psylloidea. 42
Figura 3.2 – Boxplots da contagem de Psylloidea por talhão, ao longo de coleta. . . . 43
Figura 3.3 – Boxplots da contagem de Psylloidea por altura, ao longo de coleta. . . . 43
Figura 3.4 – Gráficos de barras da contagem de Psylloidea pela combinação entre os fatores talhão e altura – Localização 1. 46
Figura 3.5 – Gráficos de barras da contagem de Psylloidea pela combinação entre os fatores talhão e altura – Localização 2. 46
Figura 3.6 – Gráficos de perfil da contagem de Psylloidea ao longo do tempo, identificados por talhão, altura e ponto – Localização 1. 47
Figura 3.7 – Gráfico de perfil individual – Localização 1 48

Figura 3.8 – Gráficos de perfil da contagem de Psylloidea ao longo do tempo, identificados por talhão, altura e ponto – Localização 2. 49
Figura 3.9 – Gráfico de perfil individual – Localização 2 50
Figura 3.10–*Worm plot* dos modelos com distribuições de probabilidade testadas. . . 53
Figura 3.11–*Worm plot* dos modelos com distribuições de probabilidade pré-selecionadas, com preditor linear para μ e diferentes estruturas aleatórias. 54
Figura 3.12–*Worm plot* dos modelos com distribuições de probabilidade pré-selecionadas, com preditor linear para μ e σ. 56
Figura 3.13–Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo I. 57
Figura 3.14–Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo I - modelo final. 60
Figura 3.15–*Worm plot* do modelo final - Binomial Negativa Tipo I. 62
Figura 3.16–*Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ. 65
Figura 3.17–*Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ. 66
Figura 3.18–Modelos com distribuição de probabilidade Binomial Negativa Tipo I e Tipo II com melhor desempenho. 67
Figura 3.19–Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II. 68
Figura 3.20–*Worm plot* do modelo final - Binomial Negativa Tipo II. 71
Figura 3.21–Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II - modelo final. 72
Figura 3.22–*Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ. 74
Figura 3.23–*Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ. 75
Figura 3.24–Modelos com distribuição de probabilidade Binomial Negativa Tipo I e Tipo II com melhor desempenho. 76
Figura 3.25–Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II. 77
Figura 3.26–*Worm plot* do modelo final - Binomial Negativa Tipo II. 79
Figura 3.27–Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II - modelo final. 80

Lista de tabelas

Tabela 2.1 – Estruturas de Matrizes de Covariância 25
Tabela 2.2 – Distribuições de probabilidade da família exponencial. 30
Tabela 2.3 – Diferentes padrões apresentados pelo *worm plot* e suas respectivas conclusões a respeito do ajuste. Adaptado de (Stasinopoulos et al., 2017, p. 428). 38
Tabela 3.1 – Medidas resumo da variável contagem de Psylloidea, por talhão. 44
Tabela 3.2 – Medidas resumo da variável contagem de Psylloidea, por altura. 45
Tabela 3.3 – Medidas resumo da variável contagem de Psylloidea, por coleta. 45
Tabela 3.4 – Identificação das unidades experimentais. 51
Tabela 3.5 – Distribuições de probabilidade testadas. 52
Tabela 3.6 – Modelos ajustados com distribuições Delaporte, Binomial Negativa Tipo I e II, com preditor linear para μ. 53
Tabela 3.7 – Teste da razão de verossimilhança entre os modelos com diferentes estruturas aleatórias. 55
Tabela 3.8 – Modelos ajustados com distribuições Delaporte, Binomial Negativa Tipo I e II, com e sem preditor linear para σ. 55
Tabela 3.9 – Medidas descritivas dos resíduos simulados do modelo Binomial Negativa Tipo I. 57
Tabela 3.10 – Estimativas dos coeficientes de regressão e parâmetros de variância do modelo Binomial Negativa Tipo I. 58
Tabela 3.11 – Modelos testados com a distribuição Binomial Negativa Tipo I. 59
Tabela 3.12 – Estimativas dos coeficientes de regressão e parâmetros de variância do modelo final - Binomial Negativa Tipo I. 61
Tabela 3.13 – Medidas descritivas dos resíduos simulados do modelo final - Binomial Negativa Tipo I. 62
Tabela 3.14 – Estimativas dos coeficientes de talhão com diferentes *baselines*. 63

Tabela 3.15–Estimativas dos coeficientes de regressão e parâmetros de variância do modelo final com talhões agrupados - Binomial Negativa Tipo I. 64
Tabela 3.16–Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ. 65
Tabela 3.17–Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ. 66
Tabela 3.18–Teste da razão de verossimilhança entre os modelos com diferentes estruturas aleatórias. 67
Tabela 3.19–Medidas descritivas dos resíduos simulados do modelo Binomial Negativa Tipo II. 67
Tabela 3.20–Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II. 69
Tabela 3.21–Modelos testados com a distribuição Binomial Negativa Tipo II. 70
Tabela 3.22–Medidas descritivas dos resíduos simulados do modelo final - Binomial Negativa Tipo II. 70
Tabela 3.23–Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II - modelo final. 73
Tabela 3.24–Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ. 73
Tabela 3.25–Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ. 74
Tabela 3.26–Teste da razão de verossimilhança entre os modelos com diferentes estruturas aleatórias. 76
Tabela 3.27–Medidas descritivas dos resíduos simulados do modelo Binomial Negativa Tipo II. 77
Tabela 3.28–Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II. 78
Tabela 3.29–Modelos testados com a distribuição Binomial Negativa Tipo II. 79
Tabela 3.30–Medidas descritivas dos resíduos simulados do modelo final - Binomial Negativa Tipo II. 79
Tabela 3.31–Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II - modelo final. 81

Sumário

Introdução **12**
1 Revisão de Literatura **15**
2 Materiais e Métodos **19**
2.1 Materiais 19
2.2 Métodos 21
2.2.1 Modelos Lineares Mistos 21
2.2.1.1 Matrizes de Covariâncias 24
2.2.1.2 Estimação por Máxima Verossimilhança 24
2.2.2 Modelos Lineares Generalizados (MLG) 26
2.2.2.1 Estimação por Máxima Verossimilhança 28
2.2.3 Modelos Lineares Generalizados Mistos (MLGM) 29
2.2.3.1 Estimação por Máxima Verossimilhança 32
2.2.4 GAMLSS - *Generalized Additive Models for Location, Scale and Shape* 33
2.2.4.1 Estimação 34
2.2.4.2 Algoritmos de Estimação RS e CG 35
2.2.4.3 Seleção de Modelos 35
2.2.4.4 Análise Diagnóstica 36
2.2.4.4.1 Resíduo Quantílico Aleatorizado 36
2.2.4.4.2 Worm plot 38
2.2.4.4.3 Função `plot()` 39
3 Aplicações **41**
3.1 Análise Descritiva 42
3.2 Modelagem 1 – Geral 49
3.3 Modelagem 2 – Localização 1 64
3.4 Modelagem 3 – Localização 2 71

4 Conclusão **83**

Referências **85**

Introdução

A agricultura é uma das principais bases da economia brasileira, sendo o agronegócio responsável por impulsionar o PIB do país e segurar a atividade econômica em momentos de recessão, uma vez que o Brasil é um dos maiores exportadores de produtos agrícolas do mundo. Neste cenário, a citricultura se destaca como uma das atividades mais relevantes do setor, dado que o Brasil é o líder na produção mundial de laranja doce e de suco concentrado de laranja. Com isso, várias pesquisas têm sido produzidas em busca de respostas em relação a melhoramentos na produção e prejuízos causados por doenças e pragas, fornecendo dados que devem ser analisados com o apoio da Estatística no intuito de subsidiar os interesses do pesquisador.

Nesse sentido, os dados gerados por pesquisas desse tipo são, geralmente, de natureza discreta ao representar contagens de organismos em uma planta e a presença ou ausência de uma determinada doença, por exemplo. Além disso, em alguns casos, tem-se o interesse em acompanhar o comportamento da resposta ao longo do tempo, caracterizando dados longitudinais e com medidas repetidas. Desta forma, a escolha da metodologia adequada para a construção de um modelo deve respeitar a natureza dos dados e o planejamento do experimento.

Destaca-se, na literatura, a utilização de modelos lineares mistos (MLM) e modelos lineares generalizados (MLG), propostos por Laird e Ware (1982) e Nelder e Wedderburn (1972), respectivamente, que permitem a introdução de correlação entre as respostas e a definição de uma distribuição de probabilidade diferente da normal. Com a evolução da pesquisa, surgem os modelos lineares generalizados mistos (MLGM) unificando características dos modelos lineares mistos, a partir da introdução dos efeitos aleatórios, e dos modelos lineares generalizados, com distribuição de probabilidade pertencente à família exponencial. Desta forma, foi possível expandir o leque de possibilidades com relação a modelagem de dados com estrutura aninhada e/ou medidas repetidas com distribuição não

normal, por exemplo.

Ainda mais recente, Rigby e Stasinopoulos (2005) propuseram uma classe ainda mais genérica do que a família exponencial de distribuições, permitindo a modelagem de dados com características diversas por meio dos parâmetros populacionais e uma função de ligação apropriada. Com isso, é possível abranger situações em que a distribuição normal ou as distribuições pertencentes a família exponencial não são adequadas para o ajuste.

Em relação à área de aplicação do experimento, uma das principais e mais devastadoras doenças que atinge a produção de citros é o Huanglongbing (HLB), também conhecida como *greening* dos citros, nativa da Ásia e, atualmente, presente em outras partes do mundo. Está relacionada com três espécies de α-proteobactérias pertencentes ao gênero *Candidatus* Liberibacter, a saber, *Candidatus* Liberibacter africanus (CLaf), asiaticus (CLas) e americanus (CLam). A transmissão da doença pode ocorrer tanto por enxertia quando por insetos vetores, que transportam a bactéria pela saliva (BOVÉ, 2006).

Os insetos vetores associados à transmissão do HLB são os psilídeos asiático (*Diaphorina citri*) e africano (*Trioza erytreae*), sendo o primeiro responsável pela transmissão de CLas na América do Norte e na Ásia, e no Brasil, pela transmissão tanto de CLas quanto de CLam. Já o psilídeo africano transmite CLaf no Oriente Médio e na África. A transmissão da bactéria pelo inseto infectado ocorre pela saliva quando este alimenta-se das folhas, enquanto que um psilídeo sadio pode adquirir o patógeno alimentando-se de uma árvore infectada (HALBERT; MANJUNATH, 2004).

Existem vários problemas relacionados ao HLB, uma vez que o controle da doença e a eliminação de plantas infectadas é de difícil execução. Alguns dos motivos para tal agravante é que os sintomas da doença demoram para se manifestar e torna-se difícil identificar as plantas realmente infectadas. Além disso, os insetos vetores percorrem tanto curtas quanto longas distâncias e, portanto, as áreas que não são controladas, como espaços abandonados e áreas vizinhas, possibilitam a propagação da doença. Não menos importante, o fator custo também torna o problema ainda maior quando a doença pode acarretar em perdas econômicas devido à devastação, e ainda, exige investimentos para o controle da doença. Algumas das possibilidades no combate ao HLB envolvem a remoção de árvores infectadas e pulverização de inseticida para controlar a propagação do psilídeo, além do plantio de mudas livres da doença (BOVÉ, 2006).

Portanto, entender mais a respeito da distribuição e o comportamento dos insetos vetores de HLB em relação ao ambiente de produção de citros é fundamental para que se tracem estratégias de controle e manejo da doença. Assim sendo, o estudo abordado no trabalho consistiu em realizar o monitoramento da altura de dispersão por voo de *D. Citri*

e, também, de outras espécies de insetos vetores presentes em pomares comerciais de laranja com a presença de plantas sintomáticas da doença. Objetiva-se encontrar modelos estatísticos com distribuição para a variável resposta contagem de insetos que forneça um bom ajuste para subsidiar os interesses do experimento em questão, com base na teoria apresentada, em particular, com enfoque em modelos lineares generalizados mistos (MLGM) e modelos aditivos para locação, escala e forma (GAMLSS). Serão levados em consideração possíveis variáveis explicativas e efeitos aleatórios que podem exercer influência sobre o comportamento da distribuição dos insetos e diferentes distribuições de probabilidade serão testadas em busca de um melhor ajuste.

Capítulo 1

Revisão de Literatura

Os modelos lineares mistos podem ser definidos como uma classe de modelos estatísticos para variáveis respostas contínuas e erros normalmente distribuídos, mas com os pressupostos de independência e homocedasticidade dos modelos clássicos de regressão sendo flexibilizados com a inclusão de efeitos aleatórios. O nome se dá pelo fato de que tais modelos são lineares nos parâmetros e compõem um misto de efeitos fixos e efeitos aleatórios (WEST; WELCH; GALECKI, 2007).

A possibilidade de incluir correlação entre as observações e modelar a variância atribui grande importância aos modelos lineares mistos devido à possibilidade de analisar dados hierárquicos, longitudinais e com medidas repetidas. Isto é, indivíduos pertencentes a um mesmo grupo, para o caso de dados hierárquicos, e medidas realizadas em um mesmo indivíduo, sob diferentes condições, não podem ser consideradas independentes e um modelo de regressão linear clássico não deve ser utilizado. Na literatura, a teoria de modelos lineares mistos foi discutida e aplicada extensivamente por vários pesquisadores tais como Pinheiro e Bates (2006), McCulloch e Searle (2001), Diggle et al. (2002), Fitzmaurice, Laird e Ware (2012), dentre outros.

West, Welch e Galecki (2007) apresentam um guia prático e didático de modelos lineares mistos com o uso de softwares estatísticos. Oliveira et al. (2021) propuseram uma aplicação de modelos lineares de efeitos mistos para avaliar a densidade básica da madeira de *Eucalyptus*, considerando diferentes espécies. Segundo os autores, a qualidade dos produtos madeireiros está associada principalmente à densidade básica, sendo esta influenciada por diversos fatores não mensurados. Deste modo, o objetivo da aplicação foi considerar a variabilidade entre espécies de árvores, entre indivíduos de uma mesma espécie ou dentro de tronco em uma mesma árvore, caracterizando correlação entre as

observações.

A teoria dos modelos lineares generalizados (MLG), proposta por Nelder e Wedderburn (1972), também surge como uma alternativa aos modelos clássicos de regressão, no sentido de que a distribuição da variável resposta deve pertencer à família exponencial de distribuições, permitindo também a modelagem de dados discretos. Dentre alguns autores que discutem a teoria estão McCullagh e Nelder (1989), Dobson (2002), Cordeiro e Demétrio (2008), Fox (2015) e Hardin e Hilbe (2018).

Na área florestal, foram desenvolvidos trabalhos como o de Hess et al. (2015), que empregou a teoria de modelos lineares generalizados na análise do crescimento de altura de *Pinus Taeda* L. devido a sua importância na exploração da capacidade produtiva do local, segundo os autores. Já na área da saúde, Tuan, Venâncio e Nascimento (2016) analisaram os efeitos da exposição a poluentes do ar no infarto agudo do miocárdio utilizando uma regressão de Poisson, enquanto que em Santos et al. (2017), um modelo binomial negativo foi utilizado no estudo da relação das variáveis climáticas na prevalência de infecção respiratória aguda, em crianças menores de dois anos em Rondonópolis/MT.

Conforme Fox (2015), a aplicação de modelos lineares foi amplamente estendida com a metodologia dos modelos lineares generalizados, possibilitando distribuições diferentes da normal para a variável resposta. Da mesma forma, os modelos lineares mistos podem ser estendidos com a metodologia dos modelos lineares generalizados mistos (MLGM), que se beneficia dos avanços metodológicos das duas classes de modelos apresentadas anteriormente, incluindo efeitos aleatórios na modelagem do problema.

Em anestesiologia veterinária, Galdino (2015) propôs a avaliação de dores abdominais em vinte e quatro equinos utilizando a metodologia dos modelos lineares generalizados mistos, com distribuição de probabilidade Bernoulli para a resposta “olhar o flanco”, um comportamento comum em equinos com dores abdominais. Os animais foram observados por cinco avaliadores em quatro momentos distintos, caracterizando dados com medidas repetidas e, portanto, adequado ao uso de MLGMs. A partir do modelo ajustado, os autores conseguiram modelar a razão de chances da resposta relativa aos efeitos principais considerados, e com isso, obtiveram resultados condizentes ao esperado pelos pesquisadores.

Oliveira e Andrade (2019) estudaram a padronização da taxa de captura na pesca de espadarte, *Xiphias gladius*, utilizando modelos lineares generalizados e modelos lineares generalizados mistos. Os dados da taxa de captura por unidade de esforço (CPUE) têm sua importância na gestão para as pescarias e, frequentemente, suas estimativas são calculadas baseadas no fator ano utilizando MLG e MLGM. Ainda, segundo os autores, o uso de MLGMs na ciência da pesca é recomendado devido aos registros de captura e esforço serem realizados em diferentes momentos mas nas mesmas embarcações, caracterizando

possível correlação entre as observações. Como resultado, concluiu-se que as estimativas sugerem que a pesca atuneira tem sido realizada de forma sustentável, em conformidade com as conclusões e recomendações do ICCAT.

Uma extensão ainda maior das classes de modelos citadas anteriormente é o GAMLSS (*Generalized Additive Models for Location, Scale and Shape*), que permite o ajuste de variáveis resposta cuja distribuição de probabilidade pertença a uma classe mais genérica em relação à família exponencial e a modelagem de outros momentos da distribuição por meio da especificação do preditor linear e de uma função de ligação apropriada. Além disso, possibilita a inclusão de efeitos aleatórios e de funções de suavização, lineares ou não lineares, no preditor linear, de tal modo que são conhecidos por serem semi-paramétricos. Na literatura, pode-se consultar Rigby e Stasinopoulos (2005), Stasinopoulos e Rigby (2008) e Stasinopoulos et al. (2017).

Serinaldi (2011) utilizou a classe de modelos aditivos generalizados para locação, escala e forma (GAMLSS), na previsão de curto prazo do preço de eletricidade, como uma alternativa aos atuais métodos estocásticos utilizados para essa finalidade. O autor comparou a abordagem utilizada com resultados obtidos por diferentes modelos e concluiu que o GAMLSS, por meio de uma estrutura considerada simples, fornece mais uma ferramenta de validação cruzada com resultados satisfatórios.

Mayr et al. (2012) aplicaram GAMLSS a dados de guia de aluguel em Munique, Alemanha, para calcular o aluguel médio em função de características espaciais e do próprio imóvel. Propuseram um algoritmo *boosting*, por meio do pacote `gamboostLSS` no R, que permite, simultaneamente, a estimativa dos efeitos do preditor e a seleção de variáveis, para contornar limitações relacionadas a alta dimensão dos dados e quantidade elevada de covariáveis.

Mais recentemente, na área ambiental, Sá, Turkman e Pereira (2018) estudaram a relação entre a incidência de incêndios e variáveis explicativas, tais como vegetação, precipitação e fatores antropogênicos, e Pitt et al. (2020) exploraram a frequência e a gravidade dos riscos catastróficos associados a variáveis consideradas como fatores de risco, tais como variáveis geográficas, metereológicas e climáticas, utilizando modelos GAMLSS.

Em particular, referente a dados agronômicos, geralmente se tem dados binominais, proporções, contagens e excesso de zeros, o que limita boa parte dos modelos que podem ser utilizados nestas circunstâncias, isto é, para dados não normais. Em alguns casos, tem-se dados longitudinais com medidas repetidas, que também necessita de uma abordagem correta para lidar com a dependência entre medidas realizadas no mesmo indivíduo. Destaca-se, nesse cenário, o uso de modelos lineares generalizados mistos, que concilia todas as características mencionadas.

Carvalho (2019) realizou uma pesquisa bibliográfica a respeito de análises estatísticas na área das Ciências Agrárias com o objetivo de atestar a relação entre a qualidade e a negligência da estatística. Posteriormente, discutiu métodos considerados mais adequados na modelagem de dados agronômicos, dentre os quais estão os modelos lineares generalizados, modelos aditivos generalizados e análise bayesiana.

Almeida et al. (2016) utilizaram um modelo binomial negativo com excesso de zeros para modelar o número de plantas infectadas por Huanglongbing em pomares comerciais de laranja, na região noroeste do Paraná. Alves (2017) utilizou a teoria dos modelos lineares generalizados com o objetivo de verificar a existência de características de variedades cítricas de copa e porta-enxerto que afetam os aspectos bioecológicos do inseto *Diaphorina Citri*, patógeno de Huanglongbing. Batista (2020) propôs os modelos COM-Poisson, Poisson generalizada restrita, *double* Poisson e contagem gama que, por meio de um parâmetro adicional para flexibilizar a análise dos dados, são alternativos aos modelos tradicionais para problemas de contagem com superdispersão e/ou subdispersão.

Também encontra-se na literatura o uso de estatística espacial ou geoestatística, que estuda o comportamento de dados com variabilidade espacial. Nesse sentido, Nava (2018) utilizou um modelo binomial com distribuição espacial em uma aplicação de infestação de *Spodoptera frugiperda* em plantação de milho. Grego (2019) utilizou a teoria dos modelos lineares generalizados geoestatísticos (MLGG), com distribuição de Poisson, na modelagem espacial da ocorrência de Huanglongbing em plantas doentes.

Nesse sentido, destaca-se grande importância atribuída aos modelos lineares generalizados e modelos lineares generalizados mistos, com diferentes abordagens sendo propostas na literatura. Com o recente avanço metodológico provocado pelos modelos aditivos generalizados para locação, escala e forma (GAMLSS), as possibilidades aumentaram significativamente. Embora haja alguns trabalhos envolvendo dados de contagem na área agronômica, notou-se uma escassez associada a dados com medidas repetidas no mesmo indivíduo. Assim, a proposta deste trabalho é realizar uma modelagem adequada considerando a estrutura dos dados, a fim de encontrar um bom ajuste e discutir os resultados a partir das metodologias apresentadas.

Capítulo 2

Materiais e Métodos

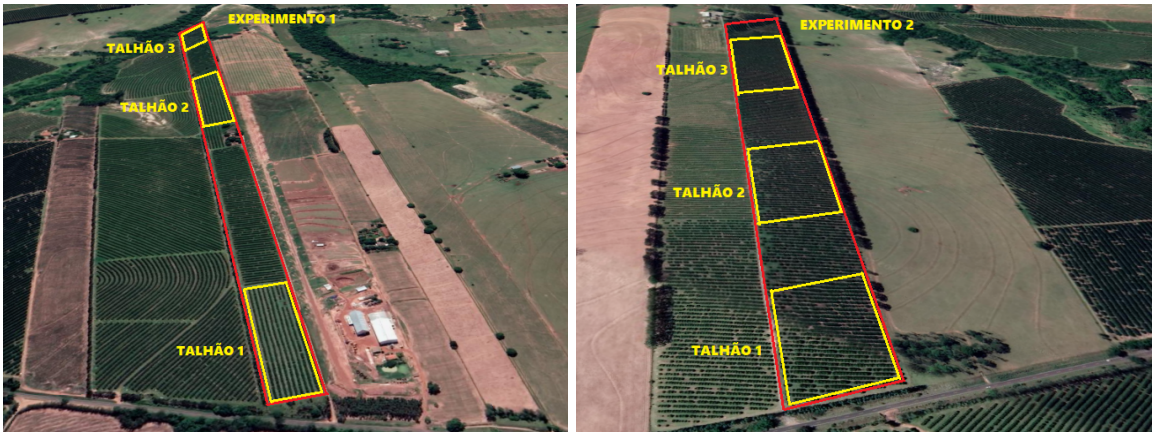
2.1 Materiais

A citricultura brasileira tem grande importância para a economia do país, tendo participação relevante na geração de empregos, balança comercial e contribuição para o PIB do país. O Brasil é o líder na produção mundial de laranja doce juntamente com Estados Unidos, China e Índia, sendo majoritariamente produzida no estado de São Paulo e seguido por Minas Gerais, Paraná, Bahia e Rio Grande do Sul, em menor relevância. Estima-se que o Brasil é responsável por mais de 30% da produção mundial de laranja, e com relação ao suco concentrado de laranja, detém mais de 50% da produção mundial sendo que quase a totalidade produzida no país é exportada. Desta forma, dada a importância da produção de citros para a atividade econômica do país e, em especial, para os produtores que dependem dessa atividade, se faz necessário o combate eficiente contra pragas e patógenos que podem potencialmente limitar ou dizimar totalmente a produção (TENNANT et al., 2009).

A condução experimental foi realizada na cidade de Paranavaí, Paraná (23°01'50"S, 52°23'18"W) pela aluna Wérica Bruna da Silva Valim em seu trabalho de mestrado em Agronomia. Foram realizados dois estudos em pomares comerciais de laranja com incidência de HLB, sendo o primeiro localizado no Sítio Nossa Senhora Aparecida, na PR258 que liga a cidade de Paranavaí a São João do Caiuá, Km 05, e o segundo, localizado a 3 km, à frente do primeiro experimento. A estrutura do planejamento foi a mesma para as duas localizações, em cada uma delas, foram selecionados três talhões (Figura 2.1) e dentro de cada talhão, inseridos bambus em quatro pontos selecionados aleatoriamente para fixação das armadilhas adesivas (VALIM, 2021).

O experimento consistiu no monitoramento do inseto vetor *Diaphorina citri*, dos insetos

Figura 2.1 – Localização das áreas experimentais.

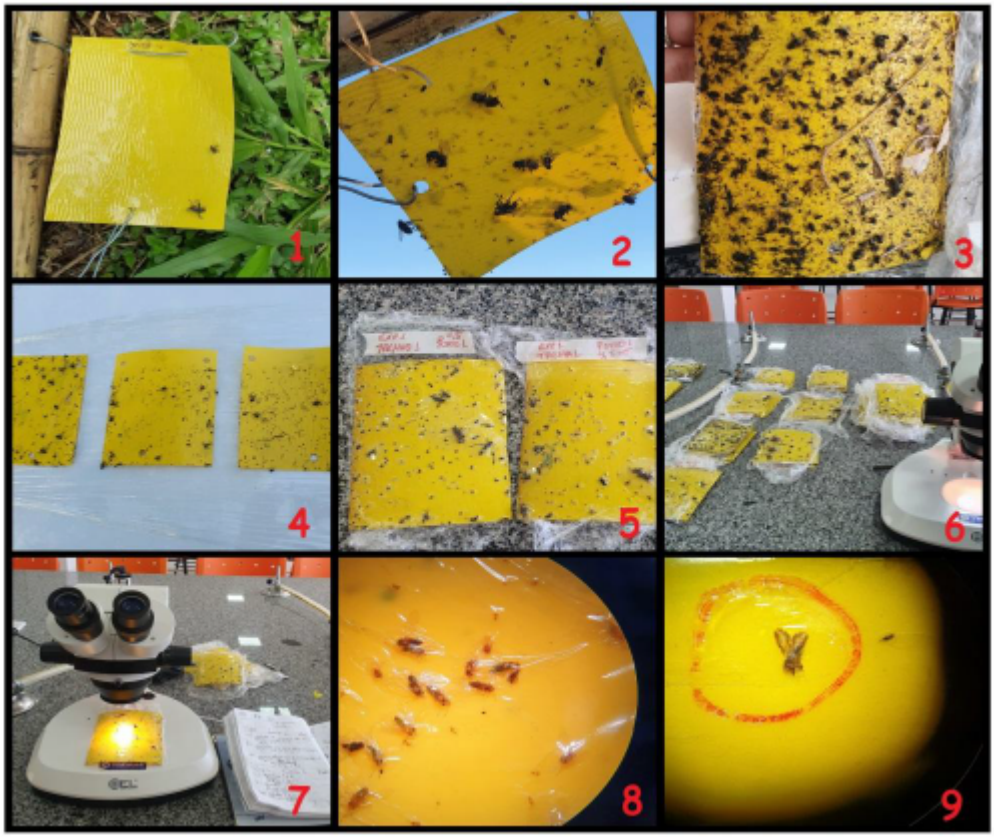


pertencentes à ordem Hemiptera: Sternorrhyncha, superfamília Psylloidea, e de tripes dos citros (*Frankliniella insularis*), utilizando armadilhas adesivas amarelas de tamanho quadrado (10cm x 10cm) fixadas em bambus perfurados em três alturas diferentes: 2,5m, 4,5m e 7m. A presença dos insetos no local provocou sua captura pelas armadilhas adesivas, permitindo, posteriormente, a identificação e contagem de cada espécie (Figura 2.2).

As armadilhas foram coletadas em períodos de 15 dias durante três meses, totalizando sete coletas, sendo etiquetadas com identificação de data, local, talhão e ponto, embaladas em sacos plásticos transparentes e refrigeradas para preservar a integridade dos insetos. Em laboratório, foram avaliadas com lupa binocular para triagem, levantamento e identificação dos vetores (VALIM, 2021).

Deste modo, tem-se um estudo longitudinal hierárquico dividido em duas localizações, no qual a contagem dos insetos foi realizada ao longo do tempo, durante três meses, em intervalos quinzenais, totalizando sete momentos diferentes para cada unidade observacional. Assim, tem-se três talhões em cada localização, quatro pontos em cada talhão e três alturas em cada ponto, observados ao longo de sete coletas, totalizando $3 \times 4 \times 3 \times 7 = 252$ observações em cada localização e 504 no banco de dados completo. Por meio do delineamento do estudo, é possível verificar a distribuição dos insetos em diferentes locais do pomar, bem como avaliar sua presença nas diferentes alturas consideradas e ao longo do tempo, e com isso, possibilitar uma assertividade maior na elaboração de um plano de ação.

Figura 2.2 – Coleta das armadilhas adesivas amarelas fixadas em pomar comercial de laranja, em que: 1, 2 e 3 – Armadilha em campo; 4 – Armadilha sequenciada para ser identificada; 5 – Armadilhas protegidas e identificadas; 6 – Separação das armadilhas por talhão e pontos; 7 – Análise em lupa binocular e anotação de contagem de insetos vetores; 8 – Armadilha vista com lupa; 9 – Identificação de *Diaphorina citri*.



2.2 Métodos

2.2.1 Modelos Lineares Mistos

Laird e Ware (1982) propõem e discutem a formulação de um modelo em dois estágios que apresenta características desejáveis para a modelagem de dados correlacionados, como ocorre em estudos longitudinais e com medidas repetidas, uma vez que as observações tomadas em um mesmo indivíduos não podem ser consideradas independentes. O modelo consiste de efeitos fixos e aleatórios e permite a modelagem da variabilidade na relação entre a resposta média e as variáveis explicativas devido à heterogeneidade da população de indivíduos. Modelos desta natureza são chamados de modelos lineares mistos, também conhecidos por modelos de efeitos mistos e modelos Laird-Ware.

Algumas das principais características desses modelos envolvem a possibilidade de

trabalhar com dados desbalanceados e modelar a variabilidade intra indivíduo e entre grupos de indivíduos, este último para dados hierárquicos. O modelo linear clássico é um caso particular da classe de modelos lineares mistos, quando contém apenas efeitos fixos no modelo.

A formulação geral dos modelos lineares mistos para o vetor de respostas $\boldsymbol{Y_i}$, de tamanho $n_i \times 1$, para o i-ésimo indivíduo, é dada na forma matricial por

$$\begin{aligned} \boldsymbol{Y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}, \quad i = 1, \cdots, N \\ \boldsymbol{b_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R_i}), \end{aligned} \tag{2.2.1}$$

isto é,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n_i} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_i1} & \dots & x_{n_ip} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1q} \\ x_{21} & x_{22} & \dots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_i1} & x_{n_i2} & \dots & x_{n_iq} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{n_i} \end{bmatrix}, \tag{2.2.2}$$

sendo N a quantidade de indivíduos, com n_i observações para o i-ésimo indivíduo. $\boldsymbol{X_i}$, de dimensão $n_i \times p$, é a matriz do modelo associada aos efeitos fixos e $\boldsymbol{\beta}$, de dimensão $p \times 1$, é o vetor de coeficientes (efeitos fixos). $\boldsymbol{Z_i}$ é uma matriz de dimensão $n_i \times q$, que pode ser um subconjunto de $\boldsymbol{X_i}$ ($q \leq p$), com as variáveis explicativas que serão modeladas como aleatórias no modelo (apenas intercepto ou inclinação aleatória, por exemplo). Os termos aleatórios são modelados pelo vetor $\boldsymbol{b_i}$, de dimensão $q \times 1$, com distribuição normal multivariada com vetor de médias zero e matriz de covariância $\boldsymbol{G}$, e por $\boldsymbol{\epsilon_i}$, que é o vetor de erros aleatórios do modelo com distribuição normal multivariada com vetor de médias zero e matriz de covariância $\boldsymbol{R_i}$, sendo que $\boldsymbol{b_i}$ e $\boldsymbol{\epsilon_i}$ são independentes entre si.

Pinheiro e Bates (2006) descrevem uma formulação básica e uma formulação geral dos modelos lineares mistos. Na formulação mais simples, os erros $\boldsymbol{\epsilon_i}$ são considerados independentes e homocedásticos, isto é, $\boldsymbol{R_i} = \sigma^2\boldsymbol{I}$. Na formulação geral estendida, a modelagem pode ser feita permitindo que os erros (intra indivíduo) sejam heteroscedásticos e correlacionados, enquanto os erros para diferentes indivíduos i são independentes entre si e independentes de $\boldsymbol{b_i}$, assim como na formulação básica.

De acordo com Costa (2003), algumas razões para se incluírem parâmetros de covariância, isto é, realizar a modelagem das matrizes $\boldsymbol{G}$ e $\boldsymbol{R_i}$, são:

- as unidades experimentais sobre as quais os dados são medidos podem ser colocadas em grupos e os dados de um grupo comum são correlacionados. Isso pode ocorrer com dados de famílias, ninhadas, colônias e pessoas que habitam a mesma casa e

- medidas repetidas são tomadas sobre a mesma unidade experimental e podem ser correlacionadas. A natureza dessas medidas pode ser multivariada. Exemplos comuns são dados observados ao longo do tempo, chamados dados longitudinais (COSTA, 2003).

Desta forma, a modelagem da variabilidade dos dados pode ser feita escolhendo-se adequadamente a estrutura das matrizes de variâncias e covariâncias dos efeitos aleatórios e dos resíduos, $\boldsymbol{G}$ e $\boldsymbol{R_i}$, respectivamente. Portanto, a partir do modelo (2.2.1), tem-se

$$\begin{aligned} E(\boldsymbol{Y_i}) &= \boldsymbol{X_i\beta} \\ Var(\boldsymbol{Y_i}) &= \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i} \end{aligned} \tag{2.2.3}$$

com as matrizes $\boldsymbol{G}$ e $\boldsymbol{R_i}$ definidas anteriormente.

Fitzmaurice, Laird e Ware (2012) ressaltam que, nos modelos lineares mistos, é necessária a distinção entre a média marginal de $\boldsymbol{Y_i}$ e a média condicional. A média condicional (ou específica por indivíduo) de $\boldsymbol{Y_i}$, dado $\boldsymbol{b_i}$, é escrita como

$$E(\boldsymbol{Y_i}|\boldsymbol{b_i}) = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}, \tag{2.2.4}$$

isto é, $\boldsymbol{b_i}$ representa o desvio da resposta média populacional, $\boldsymbol{X_i\beta}$, devido ao sujeito específico i. Podemos mostrar ainda que

$$\begin{aligned} Var(\boldsymbol{Y_i}|\boldsymbol{b_i}) &= Var(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}) \\ &= \cancelto{0}{Var(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i})} + Var(\boldsymbol{\epsilon_i}) \\ &= \boldsymbol{R_i}. \end{aligned} \tag{2.2.5}$$

Por outro lado, a média marginal (ou populacional) de $\boldsymbol{Y_i}$ é calculada quando tomamos a média sobre a distribuição dos efeitos aleatórios $\boldsymbol{b_i}$, isto é,

$$\begin{aligned} E(\boldsymbol{Y_i}) &= E(E(\boldsymbol{Y_i}|\boldsymbol{b_i})) \\ &= E(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}) \\ &= \boldsymbol{X_i\beta} + \boldsymbol{Z_i}\cancelto{0}{E(\boldsymbol{b_i})} \\ &= \boldsymbol{X_i\beta} \end{aligned} \tag{2.2.6}$$

e, além disso,

$$\begin{aligned} Var(\boldsymbol{Y_i}) &= Var(\boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}) \\ &= \cancelto{0}{Var(\boldsymbol{X_i\beta})} + Var(\boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}) \\ &= \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i}. \end{aligned} \tag{2.2.7}$$

Portanto, de acordo com os resultados apresentados acima para os modelos marginal e condicional, os modelos lineares mistos devem satisfazer

$$\boldsymbol{Y_i} \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R_i}) \tag{2.2.8}$$

e

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{b_i} &\sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{R}_i) \\ \boldsymbol{b_i} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}). \end{aligned} \tag{2.2.9}$$

2.2.1.1 Matrizes de Covariâncias

Conforme discutido anteriormente, a variabilidade da resposta pode ser modelada por meio da especificação das matrizes de covariâncias $\boldsymbol{G}$ e $\boldsymbol{R_i}$. Nos casos mais simples, a matriz de covariância para o erro aleatório $\boldsymbol{\epsilon}$ é modelada como $\boldsymbol{R}$ = $\sigma^2\boldsymbol{I}$, indicando homocedasticidade e independência entre os erros. No entanto, uma grande vantagem dos modelos lineares mistos consiste em possibilitar a especificação de uma classe de modelos mais abrangente, a partir da modelagem dessas matrizes.

O artigo de Wolfinger (1993) traz uma boa discussão em relação à modelagem da variância e inferência, sob a ótica da teoria dos modelos lineares mistos. Na Tabela 2.1, algumas das estruturas de covariâncias mais comuns são apresentadas.

Para a seleção do modelo, podem-se utilizar testes da razão de verossimilhanças, para modelos aninhados, ou critérios de informação tais como AIC (Critério de Informação de Akaike) e BIC (Critério de Informação Bayesiano) para a comparação dos modelos com diferentes estruturas.

2.2.1.2 Estimação por Máxima Verossimilhança

Um dos métodos mais comuns de estimação dos parâmetros do modelo consiste na maximização da função de verossimilhança que, considerando uma amostra aleatória de tamanho N, é dada por

$$\begin{aligned} L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \Big\{ &(2\pi)^{-n_i/2}|\boldsymbol{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \\ &exp\left[-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right]\Big\} \end{aligned} \tag{2.2.10}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, $\boldsymbol{V}_i(\boldsymbol{\alpha}) = Var(\boldsymbol{Y_i}) = \boldsymbol{ZGZ'} + \boldsymbol{R_i}$ é a matriz de variâncias e covariâncias do modelo e $\boldsymbol{\alpha}$, neste caso, representa o vetor dos parâmetros de variância das matrizes $\boldsymbol{G}$ e $\boldsymbol{R_i}$ do modelo, conhecidos por componentes de variância. De acordo com

Tabela 2.1 – Estruturas de Matrizes de Covariância

Estrutura	Notação	Matriz
Diagonal	I	$\begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \sigma^2 \end{bmatrix}$
Componente de Variância	VC(A B)	$\begin{bmatrix} \sigma_A^2 & & & \\ & \sigma_A^2 & & \\ & & \sigma_B^2 & \\ & & & \sigma_B^2 \end{bmatrix}$
Simetria Composta	CS	$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ & & \sigma^2+\sigma_1 & \sigma_1 \\ & & & \sigma^2+\sigma_1 \end{bmatrix}$
Não-estruturada	UN	$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_{33}^2 & \sigma_{34} \\ & & & \sigma_{44}^2 \end{bmatrix}$
Autoregressiva de primeira ordem	AR(1)	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$

Fonte: Wolfinger (1993).

Laird e Ware (1982), com a suposição de que $\boldsymbol{\alpha}$ é conhecido, a partir da maximização de $L(\boldsymbol{\theta})$, o estimador de máxima verossimilhança de $\boldsymbol{\beta}$ coincide com o estimador de mínimos quadrados, isto é

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{N} \boldsymbol{X_i^T W_i X_i}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X_i^T W_i y_i} \tag{2.2.11}$$

com $\boldsymbol{W_i} = \boldsymbol{V_i}^{-1}(\boldsymbol{\alpha})$.

Se $\boldsymbol{\alpha}$ for desconhecido, pode-se utilizar uma estimativa adequada $\hat{\boldsymbol{\alpha}}$ para estimar $\boldsymbol{\beta}$, utilizando $\hat{\boldsymbol{W}}_i = \boldsymbol{W}_i(\hat{\boldsymbol{\alpha}})$ em (2.2.11). Portanto, tem-se que

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{N} \boldsymbol{X_i^T} \hat{\boldsymbol{W}}_i \boldsymbol{X_i}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{X_i^T} \hat{\boldsymbol{W}}_i \boldsymbol{y_i}. \tag{2.2.12}$$

O método da máxima verossimilhança também é frequentemente utilizado para a estimação das componentes de variância, por meio da maximização da função em (2.2.10) em

relação a $\boldsymbol{\alpha}$. No entanto, segundo Laird e Ware (1982), as estimativas de $\boldsymbol{\alpha}$ são viesadas para baixo pois não levam em consideração a perda dos números de graus de liberdade na estimação dos efeitos fixos, $\boldsymbol{\beta}$. Desta forma, para corrigir o problema do viés, a estimação de $\boldsymbol{\alpha}$ é feita por meio da máxima verossimilhança restrita, uma adaptação em $L(\boldsymbol{\theta})$. Verbeke e Molenberghs (2000) mostram que a verossimilhança restrita é dada por

$$L_{REML}(\boldsymbol{\theta}) = \left|\sum_{i=1}^{N} \boldsymbol{X_i^T W_i X_i}\right|^{-\frac{1}{2}} L(\boldsymbol{\theta}), \tag{2.2.13}$$

que pode ser maximizada em relação a todos os parâmetros simultaneamente, $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$, uma vez que $\left|\sum_{i=1}^{N} \boldsymbol{X_i^T W_i X_i}\right|$ não depende de $\boldsymbol{\beta}$ (VERBEKE; MOLENBERGHS, 2000). Para mais detalhes, consultar Patterson e Thompson (1971), Harville (1974) e Verbeke e Molenberghs (2000).

Para a maximização das funções de verossimilhança é necessário empregar métodos iterativos tais como o método de Newton-Raphson e o algoritmo EM, este último descrito por Laird e Ware (1982) sob uma perspectiva bayesiana.

2.2.2 Modelos Lineares Generalizados (MLG)

Os modelos lineares generalizados (MLGs) foram propostos por Nelder e Wedderburn (1972) como uma classe de modelos cuja distribuição da variável resposta pertença à família exponencial. Além disso, a relação entre a média e a estrutura fixa do modelo pode ser modelada por uma função de ligação, monótona e diferenciável, permitindo que o preditor linear se relacione com uma transformação da média. O modelo de regressão linear clássico é um caso particular dos MLGs quando a resposta segue uma distribuição normal e a função de ligação é a identidade. Desta forma, os MLGs surgem como uma alternativa natural aos modelos tradicionais de regressão, flexibilizando as suposições de normalidade para a resposta e a relação funcional entre a média e o preditor linear.

Para uma amostra aleatória Y associada a um conjunto de variáveis explicativas $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, tem-se que a formulação dos modelos lineares generalizados segue uma estrutura de três partes, a saber, um componente aleatório, um componente sistemático e uma função de ligação. Considerando uma amostra aleatória de tamanho N, segue que:

1. **Componente aleatório:** o componente aleatório é representado por um conjunto de variáveis aleatórias $Y_1, Y_2, \ldots, Y_N$ independentes e identicamente distribuídas, cuja distribuição pertence à classe exponencial de distribuições na forma canônica. A f.d.p. de Y_i pode ser escrita como

$$f(y_i; \theta_i, \phi) = exp\{\phi^{-1}[y_i\theta_i - \psi(\theta_i)] + c(y_i, \phi)\} \tag{2.2.14}$$

sendo $E(Y_i) = \mu_i,\ i = 1, 2, \ldots, N$. Temos que θ_i é chamado de parâmetro canônico, ϕ é o parâmetro de dispersão e as funções $\psi(.)$ e $c(.)$ são conhecidas. Cordeiro e Demétrio (2008) mostram, a partir da definição da função escore, que

$$\begin{aligned} U &= \frac{\partial l}{\partial \theta} = \phi^{-1}[y - \psi'(\theta)] \\ U' &= \frac{\partial^2 l}{\partial \theta^2} = \phi^{-1}\psi''(\theta) \end{aligned} \tag{2.2.15}$$

sendo $l = \log f(y; \theta, \phi)$. Por outro lado, assumindo válidas as condições de regularidade, temos que

$$\begin{aligned} E(U) &= E\left(\frac{\partial l}{\partial \theta}\right) = 0 \\ E(U') &= E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right]. \end{aligned} \tag{2.2.16}$$

Portanto, de (2.2.15) e (2.2.16), a esperança pode ser calculada como

$$\begin{aligned} E(U) = E\{\phi^{-1}[y - \psi'(\theta)]\} &= 0 \\ \phi^{-1}[E(Y) - \psi'(\theta)] &= 0 \\ E(Y) - \psi'(\theta) &= 0 \\ E(Y) = \psi'(\theta) &= \mu. \end{aligned} \tag{2.2.17}$$

Analogamente, a partir da segunda equação em (2.2.16), podemos calcular a variância:

$$\begin{aligned} -E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = -E\{\phi^{-1}[y - \psi'(\theta)]\}^2 &= -\phi^{-1}\psi''(\theta) = E\left(\frac{\partial^2 l}{\partial \theta^2}\right) \\ -E\{\phi^{-2}[y - \psi'(\theta)]^2\} &= -\phi^{-1}\psi''(\theta) \\ \phi^{-2}E[y - \psi'(\theta)]^2 &= \phi^{-1}\psi''(\theta) \\ \phi^{-2}Var(Y) &= \phi^{-1}\psi''(\theta) \\ Var(Y) &= \phi\psi''(\theta), \end{aligned} \tag{2.2.18}$$

em que $\psi''(\theta) = V(\mu)$ é denominada função de variância. Segundo Paula (2013), a função de variância caracteriza uma classe de distribuições na família exponencial e, portanto, possui um papel importante. Algumas das distribuições pertencentes à família exponencial são Normal, Normal Inversa e Gama, para os casos contínuos, e Binomial, Binomial Negativa e Poisson nos casos discretos. A Tabela 2.2 apresenta essas distribuições e os elementos da família exponencial de cada uma.

2. **Componente sistemático:** consiste de um vetor de combinações lineares $\eta_i = \boldsymbol{x_i\beta}$, $i = 1, 2, \ldots, N$, entre as variáveis explicativas independentes $x_{i1}, x_{i2}, \ldots, x_{ip}$ e seus coeficientes $\beta_1, \beta_2, \ldots, \beta_p$, para cada Y_i, sendo $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_N)^T$ chamado de preditor linear:

$$\eta_i = \begin{bmatrix} 1 & x_{i1} & \ldots & x_{ip} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p. \tag{2.2.19}$$

3. **Função de ligação:** uma função que liga os componentes aleatório e sistemático por meio de uma transformação da média da variável resposta Y_i e sua estrutura linear (sistemática), tal que

$$\eta_i = \boldsymbol{x_i\beta} = g(\mu_i) \tag{2.2.20}$$

e g é uma função monótona e diferenciável. Quando $g(\mu_i) = \theta_i$, tem-se uma função de ligação canônica e portanto, um modelo de regressão canônico.

Portanto, a escolha de um modelo apropriado passa pela definição dessas três partes, observado o tipo de resposta da variável aleatória, uma estrutura linear e a escolha apropriada de uma função de ligação.

2.2.2.1 Estimação por Máxima Verossimilhança

A estimação dos parâmetros β pode ser feita pelo método da máxima verossimilhança. A partir da equação (2.2.14), a função de verossimilhança é definida como

$$\begin{aligned} L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i; \theta_i, \phi) &= \prod_{i=1}^{N} exp\{\phi^{-1}[y_i\theta_i - \psi(\theta_i)] + c(y_i, \phi)\} \\ &= exp\left\{\sum_{i=1}^{N} \left[\phi^{-1}[y_i\theta_i - \psi(\theta_i)] + c(y_i, \phi)\right]\right\} \end{aligned} \tag{2.2.21}$$

com $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \phi, \boldsymbol{y})$, simplificando a notação. Chamamos de *log-verossimilhança*, $l(\boldsymbol{\theta})$, o logaritmo da verossimilhança definida acima, ou seja,

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[\phi^{-1}[y_i\theta_i - \psi(\theta_i)] + c(y_i, \phi)\right]. \tag{2.2.22}$$

A estimação pelo método da verossimilhança consiste em derivar a *log-verossimilhança* em função dos parâmetros em β e igualar a zero, definindo a conhecida função *escore*, $\boldsymbol{U} = (U_1, U_2, \ldots, U_p)^T$, ou seja

$$\boldsymbol{U} = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}. \tag{2.2.23}$$

Para encontrar as equações de estimação para cada β_p, a partir da derivada em (2.2.23), se faz necessário o uso da regra da cadeia

$$U_p = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_p} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_p}. \tag{2.2.24}$$

Desta forma, definindo cada derivada separadamente, tem-se que

$$\begin{aligned}\frac{\partial l}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left\{\sum_{i=1}^{N}\left[\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\right]\right\} \\ &= \sum_{i=1}^{N}\phi^{-1}(y_i - \psi'(\theta_i)) = \sum_{i=1}^{N}\phi^{-1}(y_i - \mu_i).\end{aligned} \tag{2.2.25}$$

$$\begin{aligned}\frac{\partial \mu_i}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\{\psi'(\theta_i)\} = \psi''(\theta_i) \\ &= V(\mu_i) = V_i.\end{aligned} \tag{2.2.26}$$

$$\begin{aligned}\frac{\partial \eta_i}{\partial \beta_p} &= \frac{\partial}{\partial \beta_p}\{x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p\} \\ &= x_{ip}.\end{aligned} \tag{2.2.27}$$

A partir das equações (2.2.25), (2.2.26) e (2.2.27), pode-se definir a derivada apresentada em (2.2.24). Portanto,

$$U_p = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_p} = \sum_{i=1}^{N}\phi^{-1}(y_i - \mu_i)V_i^{-1}\frac{\partial \mu_i}{\partial \eta_i}x_{ip} \tag{2.2.28}$$

e os estimadores $\hat{\boldsymbol{\beta}}$ são encontrados a partir de $\boldsymbol{U} = \boldsymbol{0}$, para cada β_p. No entanto, as equações de estimação definidas por (2.2.28) não possuem solução analítica, sendo necessário o uso de métodos iterativos tais como Newton-Raphson e o metodo escore de Fisher.

2.2.3 Modelos Lineares Generalizados Mistos (MLGM)

Os modelos lineares generalizados mistos (MLGMs) são uma ampla classe de modelos que estendem a formulação da Seção 2.2.2 com a inclusão de efeitos aleatórios no preditor linear, de forma análoga à formulação dos modelos lineares mistos. Com as adaptações para incluir a heterogeneidade devido a efeitos não observáveis (aleatórios), isto é, dado $\boldsymbol{b_i}$, e múltiplas observações por indivíduo i, também pode ser especificado em três partes assim como os MLGs: componente aleatório, componente sistemático (preditor linear) e função de ligação. Assim, considerando uma amostra aleatória de N indivíduos e n_i observações para o i-ésimo indivíduo, tem-se que:

Tabela 2.2 – Distribuições de probabilidade da família exponencial.

Distribuição	ϕ	θ	$\psi(\theta)$	$c(y,\phi)$	$\mu(\theta)$	$V(\mu)$
Normal: $N(\mu,\sigma^2)$	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right]$	θ	1
Poisson: $P(\mu)$	1	$\log\mu$	e^θ	$-\log y!$	e^θ	μ
Binomial: $B(m,\pi)$	1	$\log\left(\frac{\mu}{m-\mu}\right)$	$m\log(1+e^\theta)$	$-\log\binom{m}{y}$	$\frac{me^\theta}{1+e^\theta}$	$\frac{\mu}{m}(m-\mu)$
Binomial Negativa: $BN(\mu,k)$	1	$\log\left(\frac{\mu}{\mu+k}\right)$	$-k\log(1-e^\theta)$	$\log\left[\frac{\Gamma(k+y)}{\Gamma(k)y!}\right]$	$\frac{ke^\theta}{1-e^\theta}$	$\mu\left(\frac{\mu}{k}+1\right)$
Gama: $G(\mu,\nu)$	ν^{-1}	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\nu\log(\nu y)-\log y-\log\Gamma(\nu)$	$-\frac{1}{\theta}$	μ^2
Normal Inversa: $IG(\mu,\sigma^2)$	σ^2	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left[\log(2\pi\sigma^2y^3)+\frac{1}{\sigma^2y}\right]$	$(-2\theta)^{-1/2}$	μ^3

Fonte: (CORDEIRO; DEMÉTRIO, 2008).

1. Dado o vetor $\boldsymbol{b_i}$, de dimensão $q \times 1$, a distribuição condicional de y_{ij}, para todo $i = 1, 2, \ldots, N$ e $j = 1, 2, \ldots, n_i$, pertence à família exponencial de distribuições com função de densidade

$$f(y_{ij}, \theta_{ij}, \phi|\boldsymbol{b_i}) = exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\} \tag{2.2.29}$$

e

$$\begin{aligned} E(Y_{ij}|\boldsymbol{b_i}) &= \mu_{ij} = \psi'(\theta_{ij}) \\ Var(Y_{ij}|\boldsymbol{b_i}) &= \phi\psi''(\theta_{ij}). \end{aligned} \tag{2.2.30}$$

2. Ao componente sistemático definido anteriormente, $\boldsymbol{\eta}$, será incluída a parte aleatória definida pelo vetor $\boldsymbol{b_i}$ de efeitos aleatórios. Tem-se que

$$\boldsymbol{\eta_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}, \tag{2.2.31}$$

sendo $\boldsymbol{\eta_i} = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{ij})^T$, $\boldsymbol{X_i}$ é a matriz do modelo para o i-ésimo indivíduo, de dimensão $n_i \times p$ e $\boldsymbol{Z_i}$ geralmente é subconjunto da matriz $\boldsymbol{X_i}$, de dimensão $n_i \times q$.

3. Por fim, a função de ligação tem a mesma forma da equação (2.2.20), relacionando o preditor linear $\boldsymbol{\eta_i}$ com uma função da média $g(\boldsymbol{\mu_i})$:

$$\boldsymbol{\eta_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} = g(\boldsymbol{\mu_i}) \tag{2.2.32}$$

com $\boldsymbol{\mu_i} = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ij})^T$.

Assim como nos modelos lineares mistos, $\boldsymbol{b_i}$ são variáveis aleatórias independentes e identicamente distribuídas, geralmente com a suposição de normalidade multivariada, $\boldsymbol{b_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$. Além disso, como também ocorre nos modelos mistos, a inclusão de efeitos aleatórios no preditor linear acarreta em correlação entre observações com efeito aleatório em comum.

Os modelos lineares mistos definidos na seção 2.2.1 são um caso particular dos MLGMs, quando a variável resposta segue distribuição normal e a função de ligação é a identidade. Segue que

1. a distribuição da resposta y_{ij}, condicional a $\boldsymbol{b_i}$, pode ser escrita como

$$\begin{aligned} f(y_{ij}, \mu_{ij}, \sigma^2|\boldsymbol{b_i}) &= \frac{1}{\sqrt{2\pi\sigma^2}}\left[-\frac{(y_{ij} - \mu_{ij})^2}{2\sigma^2}\right] \\ &= exp\left\{\frac{1}{\sigma^2}\left[y_{ij}\mu_{ij} - \frac{\mu_{ij}^2}{2}\right] - \frac{1}{2}\left[\frac{y_{ij}^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}, \end{aligned}$$

com $\phi = \sigma^2$, $\theta_{ij} = \mu_{ij}$, $\psi(\theta_{ij}) = \dfrac{\theta_{ij}^2}{2}$, $c(y, \theta_{ij}) = -\dfrac{1}{2}\left[\dfrac{y_{ij}^2}{\sigma^2} + \log(2\pi\sigma^2)\right]$ e $\mu_{ij}(\theta_{ij}) = \theta_{ij}$, conforme Tabela 2.2. Além disso, a partir da definição em (2.2.30), tem-se que

$$Var(Y_{ij}|\boldsymbol{b_i}) = \phi\psi''(\theta_{ij}) = \sigma^2.1$$

e portanto,

$$\psi''(\theta_{ij}) = V(\mu_{ij}) = 1.$$

2. o componente sistemático é definido como

$$\boldsymbol{\eta_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}$$

3. e a função de ligação é a identidade, isto é

$$\begin{aligned} g(\boldsymbol{\mu_i}) = \boldsymbol{\mu_i} &= \boldsymbol{\eta_i} \\ \boldsymbol{\mu_i} = \boldsymbol{\eta_i} &= \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} \\ E(\boldsymbol{Y_i}|\boldsymbol{b_i}) &= \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i}, \end{aligned}$$

equivalente ao definido em (2.2.4).

2.2.3.1 Estimação por Máxima Verossimilhança

A estimação dos modelos lineares generalizados mistos também baseia-se no método da máxima verossimilhança, embora sua complexidade seja maior em comparação aos modelos apresentados anteriormente. Especificando a distribuição de probabilidade conjunta de $\boldsymbol{Y_i}$ e $\boldsymbol{b_i}$ como

$$f(\boldsymbol{Y_i}, \boldsymbol{b_i}) = f(\boldsymbol{Y_i}|\boldsymbol{b_i})f(\boldsymbol{b_i}), \tag{2.2.33}$$

vimos que a distribuição de $f(\boldsymbol{Y_i}|\boldsymbol{b_i})$ pertence à família exponencial de distribuições e $f(\boldsymbol{b_i})$ tem distribuição normal. A função de verossimilhança marginal, por sua vez, pode ser determinada integrando sobre a distribuição dos efeitos não observáveis, $\boldsymbol{b_i}$, dada por

$$L(\boldsymbol{\theta}, \phi, \boldsymbol{y}) = \prod_{i=1}^{n} \int f(\boldsymbol{Y_i}|\boldsymbol{b_i})f(\boldsymbol{b_i})d\boldsymbol{b_i} \tag{2.2.34}$$

Deste modo, a maximização da função de verossimilhança em (2.2.34) em relação aos parâmetros levará às estimativas de $\boldsymbol{\beta}$, ϕ e $\boldsymbol{G}$. A complexidade neste caso se dá pelo fato de que tal maximização não possui solução analítica e, portanto, devem-se utilizar métodos de integração numérica para sua resolução tais como o método da quadratura Gaussiana e aproximação de Laplace, dentre outros.

2.2.4 GAMLSS - *Generalized Additive Models for Location, Scale and Shape*

Rigby e Stasinopoulos (2005) propuseram uma classe de modelos denominada GAMLSS (*Generalized Additive Models for Location, Scale and Shape*) que permite flexibilizar as suposições dos modelos estatísticos precedentes, tais como lidar com uma família de distribuições de probabilidade mais geral do que a família exponencial e permitir a modelagem de assimetria e curtose com a expansão da parte sistemática do modelo para outros parâmetros de uma distribuição de probabilidade paramétrica e univariada. Apesar de permitir uma modelagem mais completa e, de certa forma, mais complexa, destaca-se pela simplicidade de se utilizar as ferramentas implementadas pelo pacote `gamlss` do software **R**, ao passo que os modelos são ajustados por máxima verossimilhança por meio de algoritmos *backfitting*. Além disso, são conhecidos por serem semi-paramétricos, uma vez que exigem a definição de uma distribuição de probabilidade paramétrica mas possibilitam a inclusão de funções de suavização não-paramétricas no preditor linear.

Pela definição dada pelos autores, define-se y_i observações condicionalmente independentes em $\boldsymbol{\theta}^i$, para $i = 1, 2, \ldots, n$, com função de probabilidade $f(y_i|\boldsymbol{\theta}^i)$ e $\boldsymbol{\theta}^i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})^T$ é um vetor de p parâmetros associados às variáveis explicativas e efeitos aleatórios. Tomando $\boldsymbol{y}^T = (y_1, y_2, \ldots, y_n)$ como um vetor de observações da variável resposta cujos parâmetros $\boldsymbol{\theta}_k$, com $k = 1, 2, \ldots, p$, serão ligados às variáveis explicativas e efeitos aleatórios por meio de uma função de ligação monótona $g_k(.)$, tem-se a seguinte relação:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\gamma_{jk}, \tag{2.2.35}$$

na qual $\boldsymbol{\theta}_k$ e $\boldsymbol{\eta}_k$ são vetores de tamanho n, $\boldsymbol{\beta}_k$ é o vetor de parâmetros de regressão de tamanho J'_k, $\boldsymbol{X}_k$ e $\boldsymbol{Z}_{jk}$ são as matrizes de planejamento de dimensões $n \times J'_k$ e $n \times q_{jk}$, conforme definidas anteriormente em (2.2.1), e γ_{jk} é uma variável aleatória q_{jk}-dimensional geralmente com $\gamma_{jk} \sim N_{jk}(0, \boldsymbol{G}_{jk}^{-1})$, em que $\boldsymbol{G}_{jk}^{-1}$ é a inversa generalizada de $\boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$ e $\boldsymbol{\lambda}_{jk}$ é um vetor de hiperparâmetros, denominando o chamado modelo GAMLSS. Quando $J_k = 0$, para todo k, chega-se a um modelo completamente paramétrico dado por

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k. \tag{2.2.36}$$

Geralmente, os parâmetros populacionais $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ estão associados a locação e escala, denotados por $\boldsymbol{\mu}$ e $\boldsymbol{\sigma}$. Outros dois parâmetros, $\boldsymbol{\theta}_3 = \boldsymbol{\nu}$ e $\boldsymbol{\theta}_4 = \boldsymbol{\tau}$, podem ser definidos como parâmetros de forma e são comuns em muitas populações de distribuições de probabilidades, sendo associados a assimetria e curtose, por exemplo. Assim, considerando $k = 1, 2, \ldots, 4$, cada parâmetro $\boldsymbol{\theta}_k$ pode ser ligado ao seu preditor linear por meio de uma

função de ligação g_k, que podem ser diferentes, de tal modo que

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\gamma_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\gamma_{j2}, \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\gamma_{j3}, \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\gamma_{j4}. \end{aligned} \tag{2.2.37}$$

Devido à flexibilidade da modelagem pela função `gamlss`, é possível ajustar, por meio desta, todas as classes de modelos de regressão apresentadas neste trabalho. Selecionando, por exemplo, distribuições de probabilidade pertencentes a família exponencial e modelando apenas a média via preditor linear, com ou sem efeito aleatório, chega-se aos modelos MLG e MLGM, respectivamente. Deste modo, é possível utilizar das ferramentas fornecidas pelo pacote *gamlss* para o ajuste, comparação e seleção de modelos. Atualmente, há mais de 100 distribuições de probabilidade implementadas no pacote *gamlss.dist* que podem ser do tipo contínua, discreta ou mista.

2.2.4.1 Estimação

Para a estimação dos modelos GAMLSS, utilizam-se os métodos de máxima verossimilhança e máxima verossimilhança penalizada, para o caso de modelos paramétricos (Equação 2.2.36) e não paramétricos (Equação 2.2.35), respectivamente. A função de log-verossimilhança, dado uma amostra observada de tamanho N, é dada por

$$l = \sum_{i=1}^{N} log f(y_i|\theta^i) \tag{2.2.38}$$

enquanto que a log-verossimilhança penalizada é dada por

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \lambda_{jk}\gamma_{jk}^T \boldsymbol{G}_{jk}\gamma_{jk}. \tag{2.2.39}$$

No caso mais simples de modelos GAMLSS paramétricos, requer-se a estimação apenas de β pela maximização da função dada na Equação (2.2.38). Para os modelos GAMLSS não paramétricos, além de β, é necessário a estimação dos parâmetros γ dado um vetor de hiperparâmetros $\boldsymbol{\lambda}$ fixado, a partir da maximização de (2.2.39). A suposição de que $\boldsymbol{\lambda}$ é constante é garantida a partir da fixação desses parâmetros ou de sua estimação, que é feita automaticamente dentro dos algoritmos de estimação que serão discutidos mais a frente.

2.2.4.2 Algoritmos de Estimação RS e CG

De acordo com Rigby e Stasinopoulos (2005), os algoritmos utilizados na maximização da função de log-verossimilhança (penalizada) são generalizações dos algoritmos CG e RS, baseados em Cole e Green (1992) e Rigby e Stasinopoulos (1996), respectivamente. O algoritmo CG utiliza o valor esperado ou observado da derivada segunda, bem como derivadas cruzadas da função de log-verossimilhança em relação aos parâmetros da distribuição. No entanto, quando ocorrer dos parâmetros serem ortogonais, isto é, quando as derivadas cruzadas forem iguais a zero, a estimação por esse método não será possível. Deste modo, pode-se recorrer ao algoritmo RS que é considerado mais simples, uma vez que não utiliza derivadas cruzadas, mas pode ser computacionalmente mais lento. Ainda é possível que seja utilizado uma mistura dos dois algoritmos, que alterna entre os dois métodos durante as iterações até atingir a convergência. A definição do algoritmo a ser utilizado na função `gamlss` é dada pelo argumento `method`, que aceita as entradas `RS()`, por *default*, `CG()` e `mixed()`.

2.2.4.3 Seleção de Modelos

O processo de seleção de modelos advém da escolha de uma distribuição de probabilidade $(\mathcal{D})$, funções de ligação $(\mathcal{G})$ e dos hiperparâmetros $\boldsymbol{\lambda}$. Além disso, é necessário selecionar as variáveis/fatores que incorporarão o modelo tal como sua forma, por meio de ferramentas que possibilitem analisar sua qualidade, e também, por meio de medidas e estatísticas de teste. Desta forma, tem-se uma análise complementar que permite filtrar e selecionar um modelo considerado adequado, dado os pressupostos.

Um dos principais critérios utilizados na literatura é o Critério de Informação de Akaike (AIC - *Akaike's Information Criterion*), proposto por Akaike (1974), que é um caso particular de sua forma generalizada, o GAIC (*Generalized Akaike's Information Criterion*), dado por

$$GAIC(k) = GD + k \cdot df, \tag{2.2.40}$$

sendo $GD = -2l(\hat{\boldsymbol{\theta}})$ o Desvio Global e k é uma penalidade associada aos graus de liberdade do modelo, df. Tem-se o critério AIC tomando $k = 2$ e outros critérios, tais como o Critério de Informação Bayesiano (BIC - *Bayesian Information Criterion*), quando $k = ln(n)$, podem ser calculados. Por serem baseados no desvio global do modelo, dado um conjunto de modelos, escolhe-se aquele cujo critério tenha o menor valor. Uma vez que os critérios de informação possuem uma forma parecida e objetivam concluir a mesma coisa, optou-se por calcular apenas o AIC na comparação entre os modelos utilizando a função `GAIC()` do *gamlss*.

Para o caso de modelos aninhados, considerando dois modelos $\mathcal{M}_0$ e $\mathcal{M}_1$ com df_0 e df_1 graus de liberdade, respectivamente, obtém-se a estatística de teste

$$\Lambda = DG_0 - DG_1 \tag{2.2.41}$$

que tem distribuição assintótica χ^2 com $d = df_0 - df_1$ graus de liberdade, denominando o chamado teste da razão de verossimilhança generalizado. Para realizá-lo, pode-se utilizar a função `LR.test()` do *gamlss* (RIGBY; STASINOPOULOS, 2005).

2.2.4.4 Análise Diagnóstica

Outra etapa muito importante e indispensável para a seleção de um modelo final é a análise diagnóstica, que consiste em verificar a adequabilidade do modelo perante os pressupostos. Utiliza-se principalmente recursos gráficos para analisar o comportamento dos resíduos que, neste caso, trata-se de resíduos quantílicos aleatórizados. O *worm plot*, que será discutido adiante, é uma das principais ferramentas na análise diagnóstica pois permite verificar se o ajuste foi adequado e identificar padrões que podem levar a uma correção do modelo.

2.2.4.4.1 Resíduo Quantílico Aleatorizado

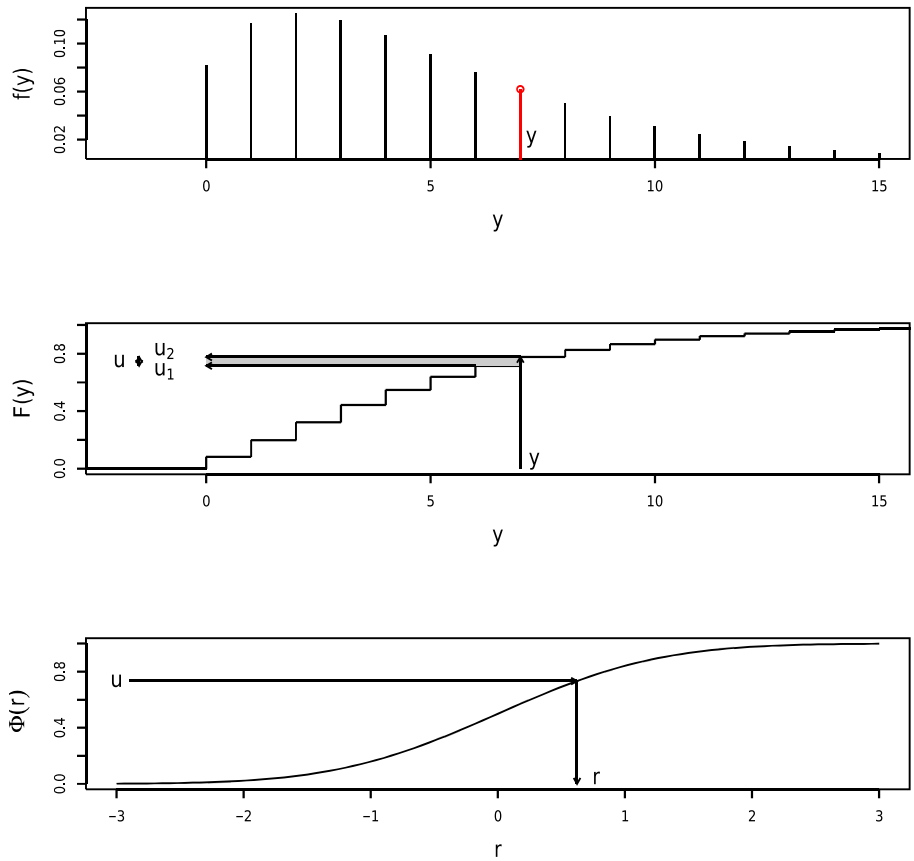
O resíduo quantílico aleatorizado, proposto por Dunn e Smyth (1996), configura uma importante ferramenta de exploração de modelos, uma vez que assegura normalidade, caso o modelo seja corretamente especificado, independente da distribuição de probabilidade da variável resposta, baseado no teorema da inversa da função de distribuição acumulada. Isso é muito útil pois o pressuposto de normalidade dos resíduos, como ocorre na regressão linear, não é satisfeito para todas as classes de modelos estatísticos devido às diferentes formas das distribuições para a resposta. É definido de formas diferentes para variáveis resposta contínuas e discretas, e neste trabalho, será apresentado apenas a forma discreta que será utilizada na aplicação.

Seja uma variável aleatória Y_i, com distribuição de probabilidade discreta e função de distribuição acumulada $F(Y_i|\boldsymbol{\theta}^i)$, tem-se que $u \sim U(u_1, u_2)$ onde

$$\begin{aligned} u_1 &= F(Y_i - 1|\boldsymbol{\theta}^i) \\ u_2 &= F(Y_i|\boldsymbol{\theta}^i). \end{aligned} \tag{2.2.42}$$

Então, o valor observado da variável aleatória, y_i, é transformado em um intervalo (u_1, u_2) associado ao "degrau" da função de distribuição acumulada, conforme exemplificado na Figura 2.3. A partir disso, é aleatorizado um valor u pertencente a esse intervalo,

Figura 2.3 – Exemplificação da transformação de uma variável observada, y, no resíduo quantílico aleatorizado, r (Stasinopoulos et al., 2017, p. 421).



que é transformado no resíduo r_i por meio da seguinte relação:

$$r_i = \Phi^{-1}(u) \tag{2.2.43}$$

em que Φ^{-1} é a inversa da função de distribuição acumulada da distribuição $N(0,1)$. Desta forma, considerando uma distribuição de probabilidade ajustada, calcula-se os resíduos $\hat{r}_i$ a partir de

$$\hat{r}_i = \Phi^{-1}(\hat{u}). \tag{2.2.44}$$

Assim, se os parametros $\boldsymbol{\theta}^i$ forem consistentemente estimados, $\hat{r}_i$ converge para uma distribuição normal padrão. Como há uma aleatorização de $\hat{u}$ no intervalo $(\hat{u}_1, \hat{u}_2)$, é importante que se façam várias simulações dos resíduos para que se tire conclusões a respeito da adequação do modelo. Para extrair os resíduos quantílicos do modelo, pode-se utilizar a função `residuals`, do pacote *stats*, a partir do objeto de saída da função `gamlss` (ajuste do modelo).

2.2.4.4.2 Worm plot

O *worm plot*, também conhecido como gráfico de minhoca, devido à sua forma, foi proposto por Buuren e Fredriks (2001) e trata-se de uma importante ferramenta de análise diagnóstica de modelos de regressão. Tal importância se atribui ao fato de que é possível atestar a adequabilidade de um modelo frente ao pressuposto de normalidade e também identificar intervalos nos quais o modelo não se ajusta adequadamente, como por exemplo, em regiões de uma variável explicativa. Segundo Stasinopoulos et al. (2017) e como pode ser visto na Tabela 2.3, é possível identificar problemas no modelo a partir de padrões visualizados nesse gráfico e, com isso, fazer correções necessárias.

Tabela 2.3 – Diferentes padrões apresentados pelo *worm plot* e suas respectivas conclusões a respeito do ajuste. Adaptado de (Stasinopoulos et al., 2017, p. 428).

Padrão do *worm plot*	Momento	Resíduo	Conclusão sobre o ajuste
Nível acima da origem	Média	Média alta	A média é subestimada
Nível abaixo da origem		Média baixa	A média é superestimada
Inclinação positiva	Variância	Variância alta	A variância é subestimada
Inclinação negativa		Variância baixa	A variância é superestimada
Concavidade para cima ($\cup$)	Assimetria	Assimetria positiva	A assimetria é baixa
Concavidade para baixo ($\cap$)		Assimetria negativa	A assimetria é alta
Forma em S com curva esquerda para baixo	Curtose	Leptocurtose	A calda é muito leve
Forma em S com curva esquerda para cima		Platicurtose	A calda é muito pesada

A construção do *worm plot* é feita a partir das diferenças entre a localização dos resíduos nas distribuições teórica e empírica. Um modelo considerado satisfatório deve ter seus pontos localizados dentro das bandas de confiança, a 95%, formado pelas duas elipses, além de não apresentar nenhuma tendência conforme descrito na Tabela 2.3. É, de certa forma, uma adaptação do *QQ plot* (gráfico quantil-quantil) removendo sua tendência. Utiliza-se a função `wp()` do *gamlss* para construí-lo, e na Figura 2.4, apresenta-se um exemplo de um *worm plot* cujo ajuste é satisfatório.

Para o caso de distribuições de probabilidade discretas, cujos resíduos simulados são obtidos por meio de um processo de aleatorização, conforme descrito anteriormente, Stasinopoulos et al. (2017) recomendam o uso da função `rqres.plot()` que realiza múltiplas realizações dos resíduos simulados e plota múltiplos gráficos *worm plot* a partir disso. O argumento `howmany` permite especificar quantas aleatorizações serão feitas, com `howmany = 6`, por *default*. Outro argumento importante é o `plot.type` que permite as entradas `"few"` e `"all"`. No primeiro, vários *worm plots* são construídos separadamentes em um único

Figura 2.4 – Exemplo de *worm plot* de um modelo bem ajustado (Stasinopoulos et al., 2017, p. 427).

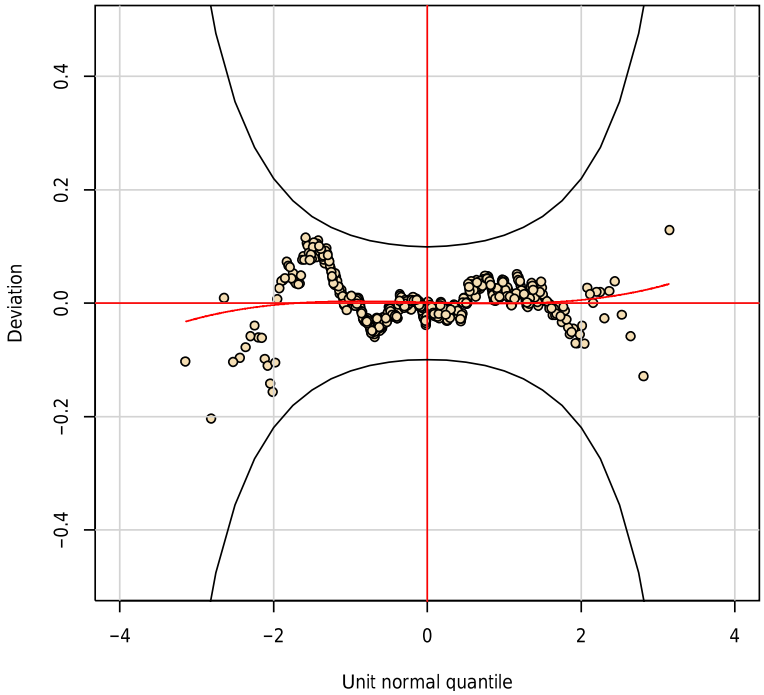


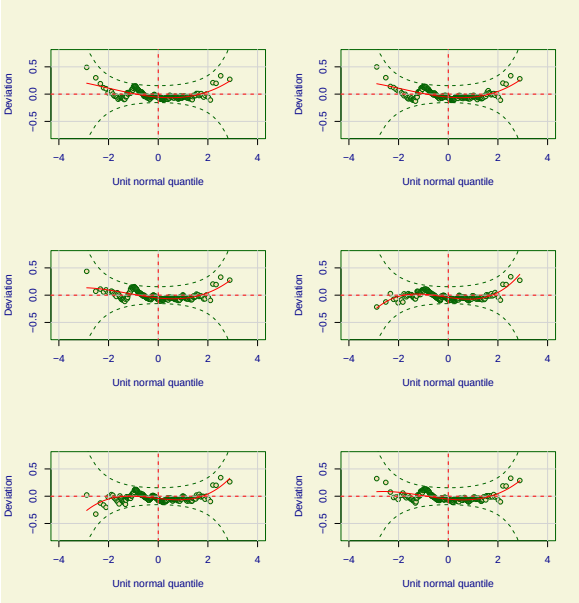
gráfico a partir das múltiplas realizações dos resíduos, que neste caso, está limitado a `howmany = 8`. Para a entrada `"all"`, as múltiplas realizações são plotadas em um único *worm plot* com as bolinhas pintadas de cinza, e em preto, traça-se a mediana dessas realizações, sem limite para `howmany`. Na Figura 2.5, apresenta-se um exemplo referente às duas situações.

2.2.4.4.3 Função `plot()`

A partir do modelo ajustado, objeto de saída da função `gamlss()`, a função `plot.gamlss()` fornece um quadro de quatro gráficos utilizando os resíduos quantílicos aleatorizados, sendo eles resíduos *versus* valores ajustados, resíduos *versus index* ou uma variável explicativa específica, gráfico de densidade estimada (*density plot*) e um *QQ plot* (gráfico quantil-quantil). Além disso, como uma das saídas da função, mostra uma tabela com medidas descritivas dos resíduos. Esse recurso gráfico permite verificar o comportamento dos resíduos quanto a normalidade e identificar se há algum padrão nos resíduos, quando espera-se que haja um comportamento aleatório em torno de zero. Também é possível utilizar a função `plot()` do pacote *base*, desde que a entrada para essa função seja um objeto da função `gamlss()`. Na Figura 2.6, exemplifica-se o uso da função `plot()`

Figura 2.5 – Exemplo de *worm plot* pela função `rqres.plot()` para distribuições de probabilidade discretas.

(a) `plot.type = "few"`



(b) `plot.type = "all"`

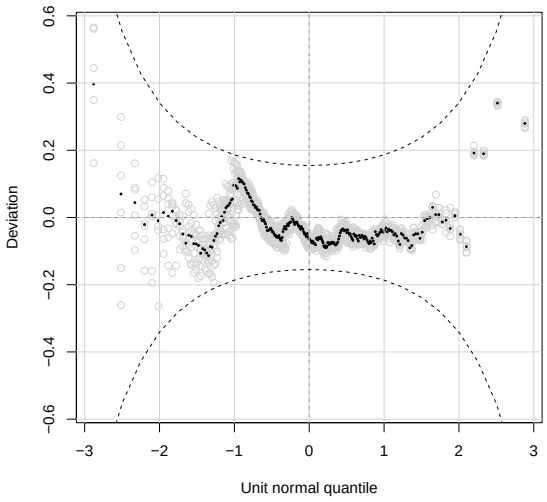
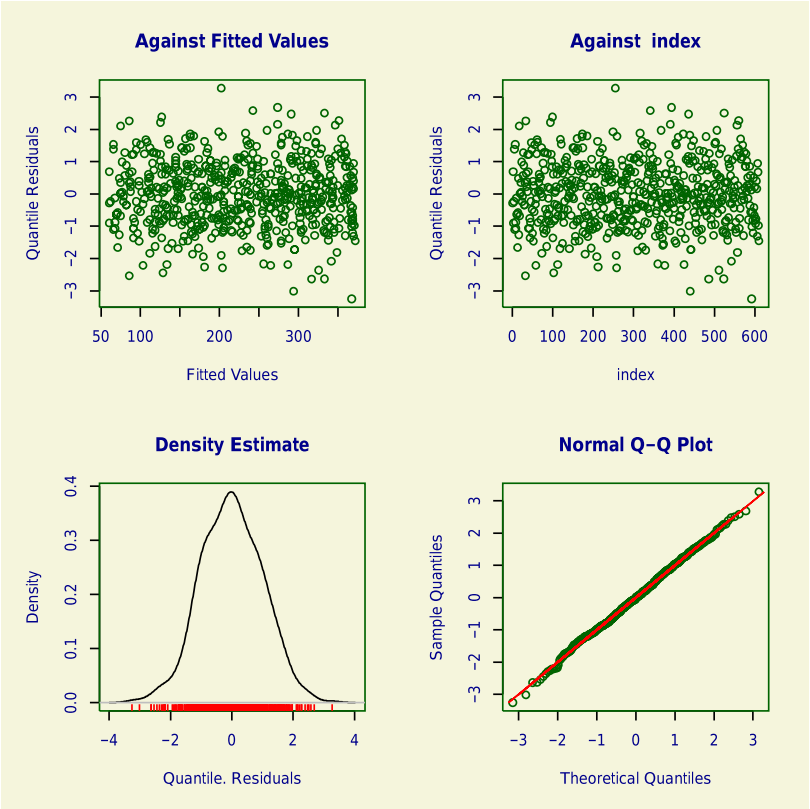


Figura 2.6 – Exemplo da função `plot()` de um modelo GAMLSS (Stasinopoulos et al., 2017, p. 424).



Capítulo 3

Aplicações

Para a modelagem de dados discretos, em particular, dados de contagem, limita-se a escolha de distribuições de probabilidade que tenham essa característica, tais como Poisson e Binomial Negativa, no caso da família exponencial de distribuições. Além disso, a função de ligação também deve ter o suporte necessário, uma vez que, para a média, a contagem deve ser maior ou igual a zero. Para o experimento abordado, tem-se a contagem de Psylloidea nas duas localizações descritas na Seção 2.1, totalizando, deste modo, três situações diferentes para análise: o banco de dados completo, contendo as duas localizações, e de forma separada.

A variável resposta considerada trata-se da contagem de insetos da superfamília Psylloidea considerando como fatores de efeitos fixos talhão, altura e coleta. Os pontos de coleta estão sendo considerados como um fator de efeito aleatório na modelagem, uma vez que representam apenas uma amostra de todas as possíveis localizações dentro do talhão. Deste modo, o interesse está na diferença entre os níveis dos fatores, talhão e altura, do efeito do tempo na distribuição da contagem de insetos, além da variabilidade dentro de talhão, representada pelo efeito aleatório de ponto. Além disso, o fator altura também pode ser incluído como fator de efeito aleatório, uma vez que dentro de um mesmo ponto, há respostas em diferentes alturas e possivelmente haja uma correlação entre as medidas.

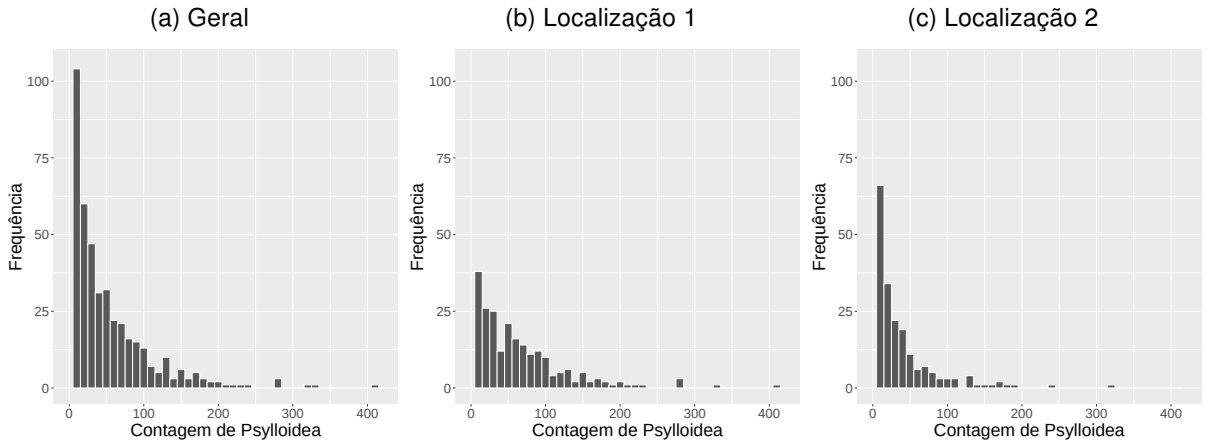
Inicialmente, serão utilizadas ferramentas de análise descritiva para verificar o comportamento da distribuição da contagem de insetos ao longo do tempo. Dentre as ferramentas utilizadas estão os gráficos comumente usados, tais como gráficos de barras, boxplots e gráficos de perfil, além de tabelas com medidas resumo, de posição e dispersão. Com isso, será possível ter uma noção inicial das principais características do banco de dados e

possíveis influências de pontos discrepantes que podem prejudicar a análise, além de verificar o comportamento da resposta associado aos níveis dos fatores. Por fim, os ajustes serão realizados por meio da função `gamlss` do pacote *gamlss* via *Software R versão 3.6.0* e, posteriormente, testados e comparados com base no critério AIC e no teste da razão de verossimilhança, além da análise de diagnóstico para verificação da qualidade do ajuste e possibilitando uma discussão a respeito dos resultados.

3.1 Análise Descritiva

Nos gráficos de barras ilustrados na Figura 3.1, é possível ter uma noção do comportamento marginal dos dados sem considerar os efeitos dos fatores do experimento, como uma primeira noção do possível comportamento da distribuição dos dados. Repara-se que a contagem dos insetos apresenta um comportamento diferente em cada localização, sendo que, na Figura 3.1b), a distribuição da contagem é mais dispersa e, na Figura 3.1c), há uma concentração até 100 com apenas alguns pontos acima dessa faixa. Na Figura 3.1a), por sua vez, tem-se que o comportamento dos dados é um meio termo entre as localizações 1 e 2.

Figura 3.1 – Gráficos de barras da distribuição de contagem de Psylloidea.



Outro gráfico muito comum em análise estatística é o boxplot, que permite visualizar de forma gráfica as principais características da distribuição dos dados, como os quartis, a dispersão e principalmente, a visualização de outliers. Além disso, um dos interesses da pesquisa consiste em verificar o comportamento da distribuição de insetos ao longo do tempo. Deste modo, nas Figuras 3.2 e 3.3, têm-se os boxplots da contagem de insetos pela ordem de coleta, por talhão e altura.

Figura 3.2 – Boxplots da contagem de Psylloidea por talhão, ao longo de coleta.

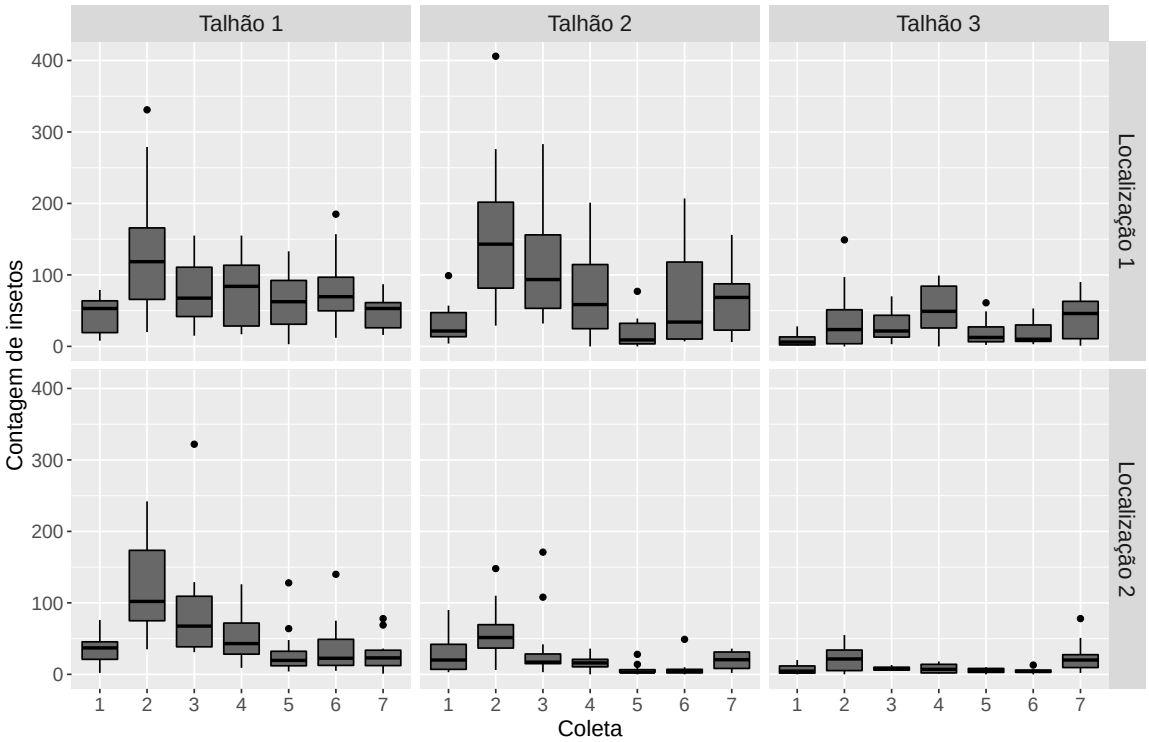
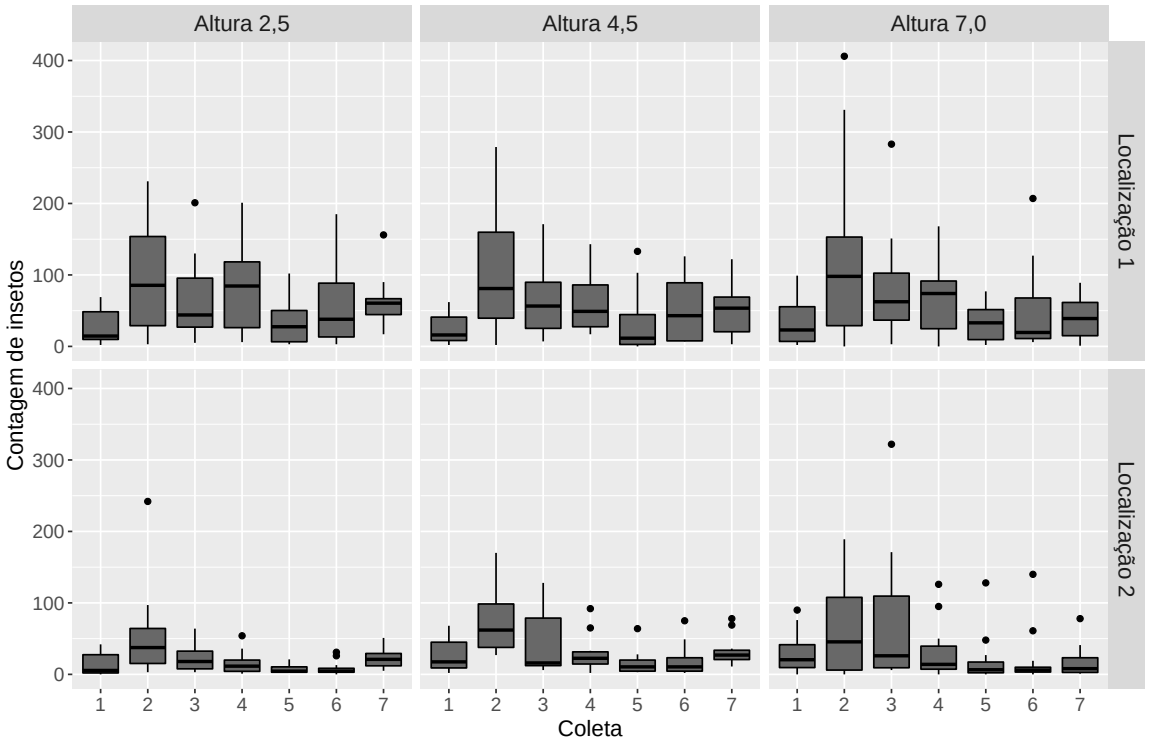


Figura 3.3 – Boxplots da contagem de Psylloidea por altura, ao longo de coleta.



Em todos esses gráficos, é possível observar que o tempo de coleta exerce influência sobre a contagem média de insetos, que oscila ao longo das coletas. Olhando para o

efeito de talhão, na localização 1, o talhão 1 parece não apresentar muita diferença na distribuição da contagem de insetos ao longo das coletas, com a mediana oscilando entre 50 e 80, com exceção da coleta 2, que tem mediana acima de 100, assim como também ocorre para a dispersão. Em relação aos talhões 2 e 3, essa diferença na contagem ao longo das coletas é mais evidente, tanto na contagem média quanto na dispersão, com padrões de crescimento e descrescimento. Comparando os talhões entre si, nota-se que apenas o talhão 3 se difere bastante dos demais, com as observações concentradas, em sua maioria, na faixa de 0 a 50.

Para a localização 2, apenas o talhão 1 se destaca dos demais com dispersão maior na contagem ao longo das coletas, enquanto talhões 2 e 3 apresentam uma distribuição extremamente concentrada. Na comparação entre as duas localizações, o que se pode concluir é que o comportamento das contagens ao longo das coletas é similar nas duas localizações, o que pode indicar que o efeito do tempo independe da localização. Já com relação à distribuição das contagens e efeito de talhão, estes apresentam características próprias em cada localização.

Observa-se, pela Figura 3.3, na localização 1, que parece não haver efeito de altura, uma vez que a distribuição da contagem de insetos ao longo de coleta é muito parecida nos tres níveis de altura. Na localização 2, parece haver alguma diferença entre os níveis, principalmente nas coletas 2 e 3, que na altura 7m, se destoa bastante do restante das coletas. Na comparação entre as localizações, de forma análoga ao que ocorre para o fator talhão, tem-se um comportamento muito parecido ao longo do tempo mas com diferenças entre os níveis de altura para cada localização, sendo que na localização 2, parece haver algum efeito. É possível observar a presença de vários outliers nas duas localizações, atingindo o valor máximo na faixa de 400, para a localização 1, e na faixa de 300 para a localização 2. Nas Tabelas 3.1, 3.2 e 3.3 apresentam-se as medidas descritivas associadas aos fatores para cada localização, considerando "Geral" como o banco de dados completo, sendo os talhões 1, 2 e 3 pertencentes à localização 1 e os talhões 4, 5 e 6 pertencentes à localização 2.

Tabela 3.1 – Medidas resumo da variável contagem de Psylloidea, por talhão.

			Mínimo	Máximo	Mediana	Média	Desvio P.
Geral	Localização 1	Talhão 1	3,00	331,00	64,00	74,11	58,43
		Talhão 2	0,00	406,00	48,00	75,68	76,81
		Talhão 3	0,00	149,00	19,50	29,39	29,76
	Localização 2	Talhão 4	1,00	322,00	37,00	57,05	56,69
		Talhão 5	0,00	171,00	15,00	24,92	31,60
		Talhão 6	0,00	78,00	6,50	11,32	13,48

Tabela 3.2 – Medidas resumo da variável contagem de Psylloidea, por altura.

		Mínimo	Máximo	Mediana	Média	Desvio P.
Geral	Altura 2,5	0,00	242,00	23,00	41,03	49,95
	Altura 4,5	0,00	279,00	30,00	46,11	48,92
	Altura 7,0	0,00	406,00	22,50	49,09	64,80
Localização 1	Altura 2,5	2,00	231,00	46,00	61,08	57,05
	Altura 4,5	0,00	279,00	42,50	56,87	57,02
	Altura 7,0	0,00	406,00	45,00	61,23	71,27
Localização 2	Altura 2,5	0,00	242,00	12,50	20,98	30,87
	Altura 4,5	2,00	170,00	22,00	35,36	36,47
	Altura 7,0	0,00	322,00	13,50	36,95	55,41

Tabela 3.3 – Medidas resumo da variável contagem de Psylloidea, por coleta.

		Mínimo	Máximo	Mediana	Média	Desvio P.
Geral	Coleta 1	0,00	99,00	18,50	26,28	24,95
	Coleta 2	0,00	406,00	56,00	88,53	85,61
	Coleta 3	3,00	322,00	38,00	59,06	63,93
	Coleta 4	0,00	201,00	29,00	46,81	46,27
	Coleta 5	0,00	133,00	10,00	24,31	31,37
	Coleta 6	0,00	207,00	11,50	35,68	47,59
	Coleta 7	0,00	406,00	22,50	49,09	64,80
Localização 1	Coleta 1	2,00	99,00	18,00	28,50	26,34
	Coleta 2	0,00	406,00	94,50	110,47	100,91
	Coleta 3	3,00	283,00	52,50	71,75	62,22
	Coleta 4	0,00	201,00	71,00	68,08	50,71
	Coleta 5	0,00	133,00	27,00	34,03	35,25
	Coleta 6	3,00	207,00	31,00	54,75	55,53
	Coleta 7	1,00	156,00	53,00	50,50	34,50
Localização 2	Coleta 1	0,00	90,00	18,50	24,06	23,64
	Coleta 2	0,00	406,00	56,00	88,53	85,61
	Coleta 3	3,00	322,00	38,00	59,06	63,93
	Coleta 4	0,00	201,00	29,00	46,81	46,27
	Coleta 5	0,00	133,00	10,00	24,31	31,37
	Coleta 6	0,00	207,00	11,50	35,68	47,59
	Coleta 7	1,00	78,00	22,00	23,94	19,87

Nas Figuras 3.4 e 3.5, apresentam-se os gráficos de barras com a combinação dos fatores talhão e altura. Na Figura 3.4, nota-se que, em cada talhão, a contagem de insetos se distribui de forma parecida entre os níveis de altura. A diferença está entre os níveis de talhão, sendo que os talhões 1 e 2 são muito parecidos e o talhão 3 é mais concentrado,

corroborando a análise do boxplot na Figura 3.2. Na Figura 3.5, parece haver algumas particularidades nesses gráficos, sendo que no talhão 1 e altura 2,5m, a contagem de insetos se concentra na faixa de 0 a 100; na altura 4,5m, a contagem de insetos é mais dispersa com frequências menores e distribuída até a faixa de 150; e na altura 7m, a contagem parece ser mais dispersa ainda e distribuída até a faixa de 200. Nos talhões 2 e 3, essa diferença entre os níveis de altura não é tão evidente, e na comparação entre os talhões, também não há diferenças muito grandes, com exceção do talhão 1 que tem uma distribuição mais espalhada.

Figura 3.4 – Gráficos de barras da contagem de Psylloidea pela combinação entre os fatores talhão e altura – Localização 1.

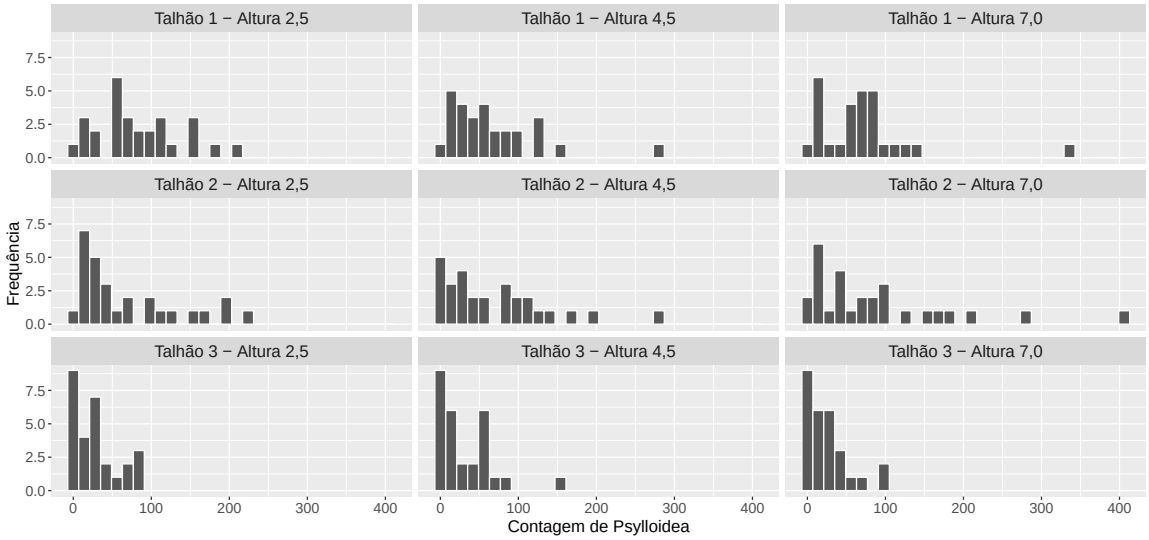
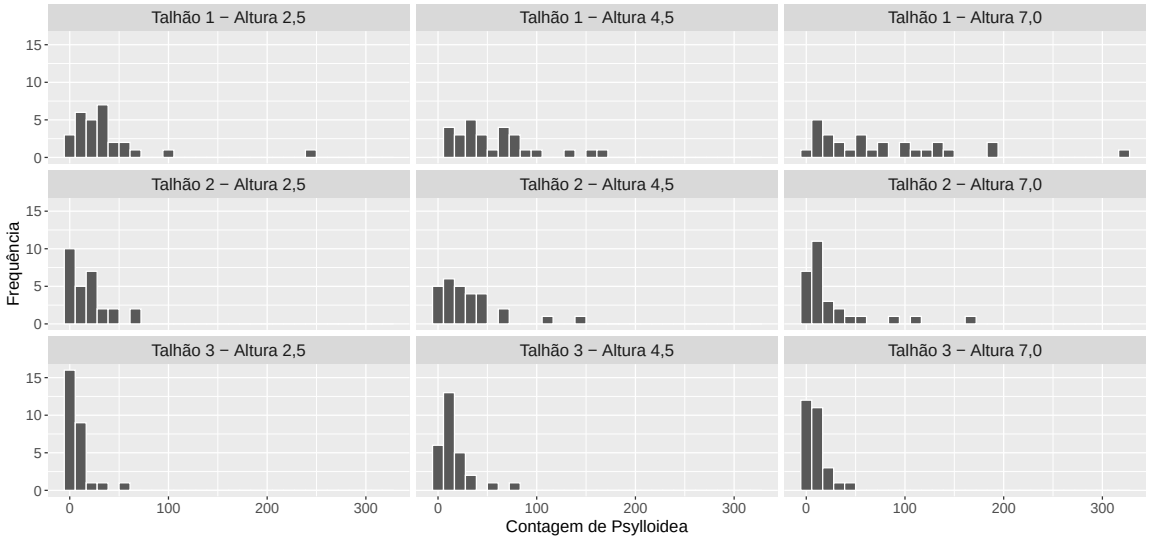
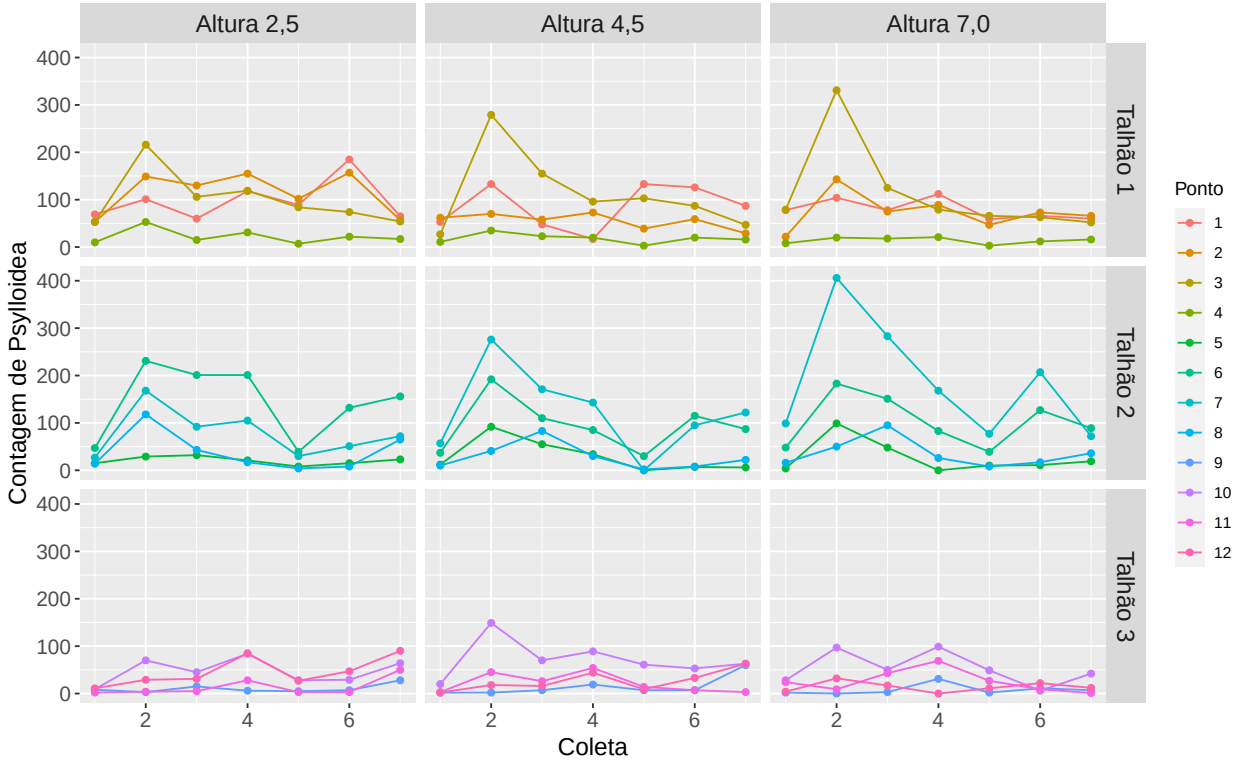


Figura 3.5 – Gráficos de barras da contagem de Psylloidea pela combinação entre os fatores talhão e altura – Localização 2.



Por fim, para complementar a análise, apresentam-se gráficos de perfil que permitem visualizar o comportamento da distribuição ao longo do tempo de forma mais individual, por meio da plotagem dos pontos, como em um gráfico de dispersão, e unindo tais pontos por segmentos de reta. Nas Figuras 3.6 e 3.8, estão identificados talhão, altura e ponto, enquanto nas Figuras 3.7 e 3.9, tem-se os gráficos de perfil para cada sujeito. Os sujeitos, identificados de 1 a 72, se referem às unidades observacionais e foram classificados de acordo com as combinações dos fatores do estudo, conforme Tabela 3.4.

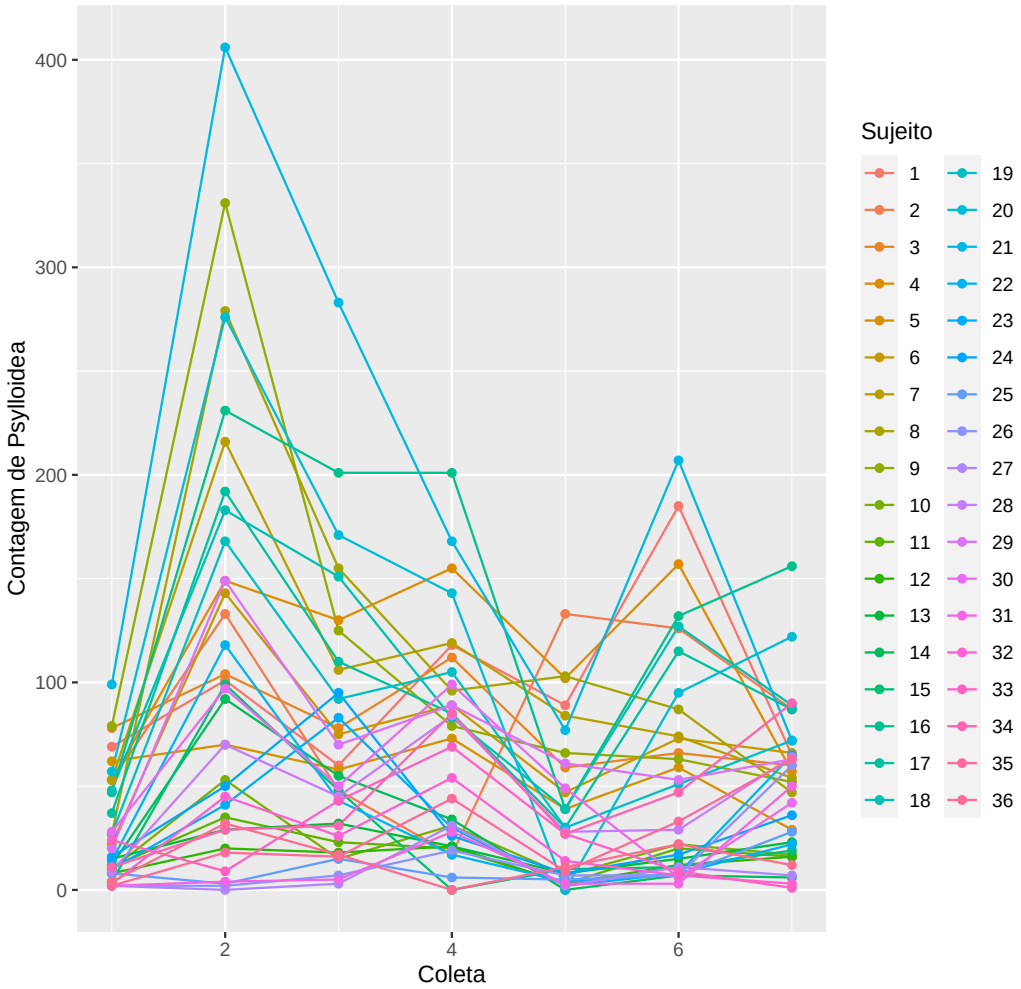
Figura 3.6 – Gráficos de perfil da contagem de Psylloidea ao longo do tempo, identificados por talhão, altura e ponto – Localização 1.



No gráfico de perfil da localização 1, vê-se que há, em alguns momentos, um crescimento da contagem nas coletas iniciais com um pico na coleta 2 e, depois, um decrescimento até a coleta 5 e novamente, voltando a crescer nas coletas 6 e 7. Na altura 7m do talhão 1, no entanto, há apenas o pico na coleta 2 e depois um padrão de constância ao longo das coletas. No talhão 3, é possível ver que há também um pico na coleta 4, além da coleta 2, voltando a crescer também nas coletas 6 e 7, configurando um claro padrão cíclico de crescimento e decrescimento. Na localização 2, por sua vez, é possível observar que há picos nas coletas 2 e 3, mas no geral, a contagem não varia muito ao longo das coletas em comparação com a localização 1.

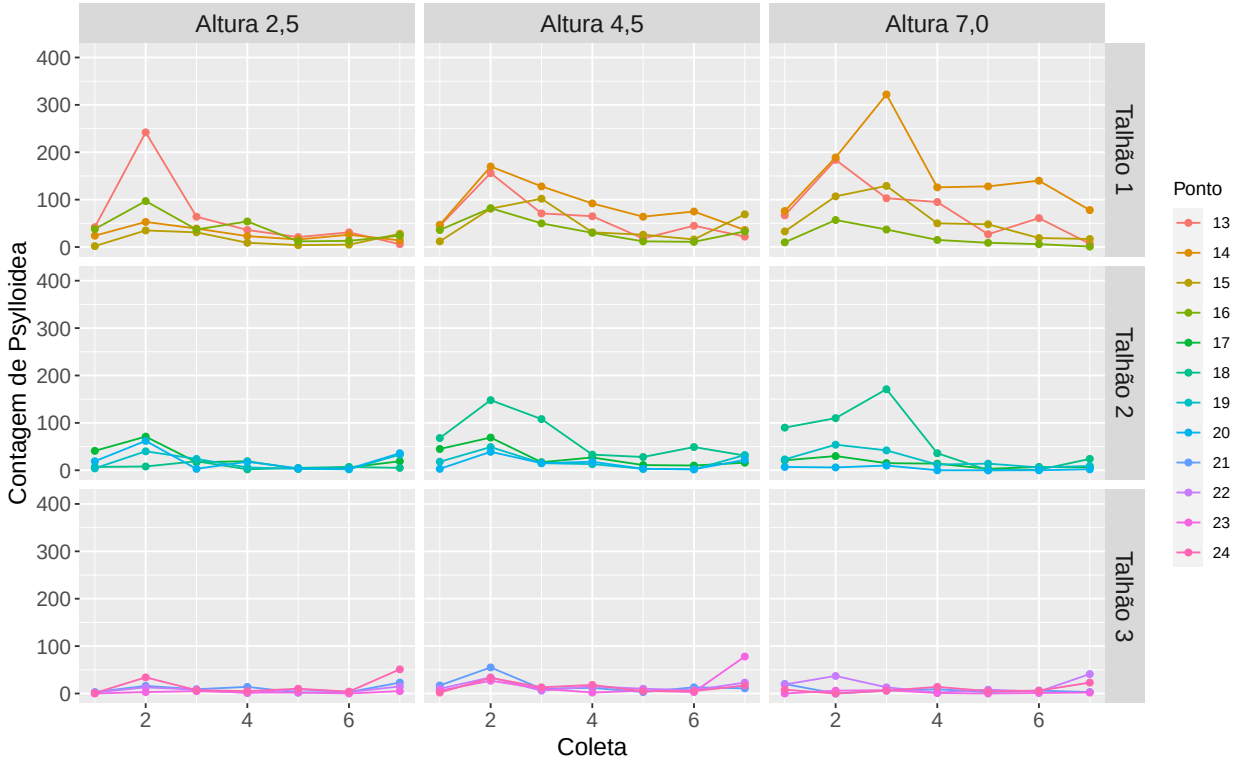
Em um panorama geral, constata-se que cada localização tem suas particularidades,

Figura 3.7 – Gráfico de perfil individual – Localização 1



levando-se em consideração o efeito dos fatores envolvidos. A localização 1 possui contagens maiores e com uma dispersão maior, oscilando bastante ao longo das coletas. Além disso, parece que o fator altura não é significativo na explicação da resposta e o para o fator talhão, apenas o talhão 3 se diferencia dos demais com uma concentração maior da contagem. A localização 2 apresenta uma distribuição da contagem de insetos mais concentrada, menos dispersa e com contagens menores. Os efeitos dos fatores talhão, altura e coleta não são tão evidentes quando para a localização 1, com poucas diferenças entre os níveis. Há picos de contagens nas coletas 2 e 3 nas quais foi possível constatar, pelos gráficos boxplot e pela Tabela 3.3, uma contagem média maior e mais dispersa em comparação com as demais coletas.

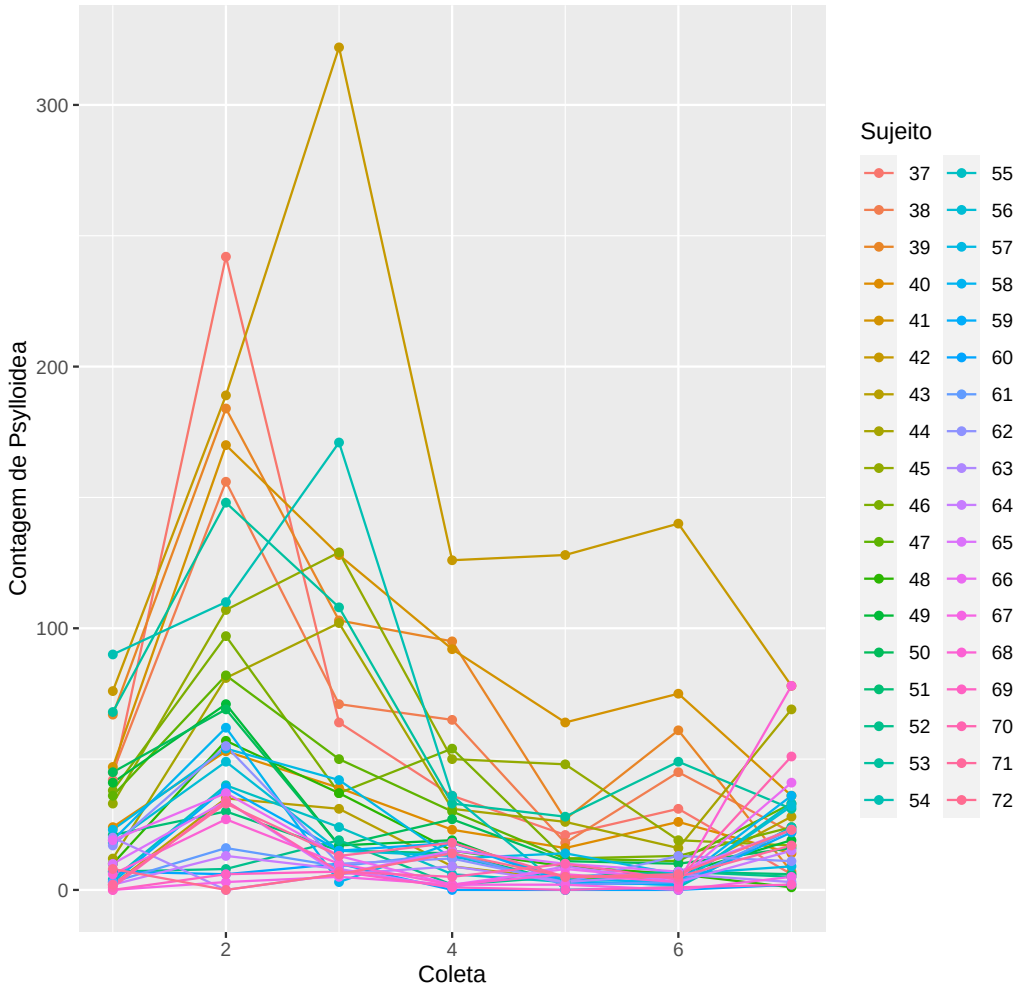
Figura 3.8 – Gráficos de perfil da contagem de Psylloidea ao longo do tempo, identificados por talhão, altura e ponto – Localização 2.



3.2 Modelagem 1 – Geral

Considerando os dados das duas localizações conjuntamente, utilizando a metodologia dos modelos GAMLSS (*Generalized additive model for location, scale and shape*), foram testadas diferentes distribuições de probabilidade buscando uma melhor aderência. Atualmente, diversas distribuições de probabilidade com suporte para variáveis de contagem estão implementadas pela função `gamlss.family`, do pacote `gamlss.dist`. Dentre elas, para a aplicação em questão, foram testadas as distribuições com dois parâmetros Poisson Generalizada, Poisson Normal Inversa, Binomial Negativa Tipo I e Binomial Negativa Tipo II, e com três parâmetros, Família Binomial Negativa, Beta Binomial Negativa, Sichel e Delaporte, conforme Tabela 3.5. Para todas elas, inicialmente, apenas a média foi modelada via preditor linear e foram considerados ponto e altura como fatores de efeitos aleatórios e talhão, altura e coleta como fatores de efeitos fixos. Posteriormente, objetivou-se a comparação entre os efeitos de talhão, por meio de contrastes, que pertencem a localizações diferentes, sendo que os talhões 1, 2 e 3 são referentes a localização 1 e os talhões 4, 5 e

Figura 3.9 – Gráfico de perfil individual – Localização 2



6 são referentes a localização 2. Portanto, tem-se

$$\begin{aligned} g_1(\mu_{ijkt}) = \eta_{ijkt} &= \beta_0 + \alpha_i + \theta_j + \gamma_t + b_{(j|k)} + b_k \\ &= (\beta_0 + b_{(j|k)} + b_k) + \alpha_i + \theta_j + \gamma_t, \end{aligned} \tag{3.2.1}$$

sendo β_0 o intercepto do modelo; α_i o efeito fixo associado ao i-ésimo talhão, $i = 1, 2, \ldots, 6$; θ_j o efeito fixo associado a j-ésima altura, $j = 1, 2, 3$ e γ_t o efeito fixo associado a t-ésima coleta, $t = 1, 2, \ldots, 7$. b_k e $b_{(j|k)}$ são os efeitos aleatórios associados ao k-ésimo ponto e j-ésima altura dentro do k-ésimo ponto, respectivamente.

Para a comparação entre as distribuições, além do critério de informação de Akaike (AIC) como medida de qualidade do ajuste, apresentado na última coluna da Tabela 3.5, foram inspecionados os gráficos *worm plots* de cada modelo para verificar sua aderência, a partir de múltiplas realizações dos resíduos simulados utilizando a função `rqres.plot()`, conforme Figura 3.10. Pode-se observar que, pelo AIC, os três melhores modelos são Família Binomial Negativa (ID 5), Delaporte (ID 8) e Binomial Negativa Tipo I (ID 3), res-

Tabela 3.4 – Identificação das unidades experimentais.

			Altura		
Localização	Talhão	Ponto	2,5	4,5	7
1	1	1	S1	S2	S3
		2	S4	S5	S6
		3	S7	S8	S9
		4	S10	S11	S12
	2	1	S13	S14	S15
		2	S16	S17	S18
		3	S19	S20	S21
		4	S22	S23	S24
	3	1	S25	S26	S27
		2	S28	S29	S30
		3	S31	S32	S33
		4	S34	S35	S36
2	1	1	S37	S38	S39
		2	S40	S41	S42
		3	S43	S44	S45
		4	S46	S47	S48
	2	1	S49	S50	S51
		2	S52	S53	S54
		3	S55	S56	S57
		4	S58	S59	S60
	3	1	S61	S62	S63
		2	S64	S65	S66
		3	S67	S68	S69
		4	S70	S71	S72

pectivamente, no entanto, o *worm plot* do modelo pela Família Binomial Negativa, Figura 3.10e), apresentou uma tendência linear negativa indicando problemas com a variância, de acordo com a interpretação das diferentes tendências apresentadas por essa ferramenta dada por Stasinopoulos et al. (2017). Como as distribuições Binomial Negativa Tipo I e Tipo II são casos particulares da Família Binomial Negativa e, além disso, apresentaram um comportamento melhor dos resíduos, foram selecionadas como candidatas potenciais para o ajuste as distribuições Delaporte, Binomial Negativa Tipo I e Binomial Negativa Tipo II. Ainda, foram testados os modelos apresentados na Tabela 3.6, pelo teste da razão de verossimilhança, com o objetivo de comparar as estruturas aleatórias e selecionar o modelo mais adequado. Os respectivos *worm plots* estão apresentados na Figura 3.11.

Tabela 3.5 – Distribuições de probabilidade testadas.

ID	Distribuição	GAMLSS	$E(Y)$	$Var(Y)$	AIC
1	Poisson Generalizada	GPO(μ, σ)	μ	$\mu(1+\sigma\mu)^2$	4567,867
2	Poisson Normal Inversa	PIG(μ, σ)	μ	$\mu+\sigma\mu^2$	4511,345
3	Binomial Negativa Tipo I	NBI(μ, σ)	μ	$\mu+\sigma\mu^2$	**4473,042 (**)**
4	Binomial Negativa Tipo II	NBII(μ, σ)	μ	$\mu+\sigma\mu$	**4497,592 (***)**
5	Família Binomial Negativa	NBF(μ, σ, ν)	μ	$\mu+\sigma\mu^\nu$	4457,693
6	Beta Binomial Negativa	BNB(μ, σ, ν)	μ	$\mu(1+\mu\nu)(1+\sigma\nu^{-1})(1-\sigma)^{-1}$, se $\sigma<1$ ∞, caso contrário.	4485,855
7	Sichel	SI(μ, σ, ν)	$b\mu$	$b\mu+b^2\mu^2 g_1$	4503,502
8	Delaporte	DEL(μ, σ, ν)	μ	$\mu+\sigma(1-\nu)^2\mu^2$	**4472,505 (*)**

(***) – Marcados com asteriscos os melhores modelos em termos de AIC e análise do *worm plot*.
A quantidade de asteriscos indica a ordem (primeiro, segundo e terceiro melhor modelo).

Fonte: os autores.

Figura 3.10 – *Worm plot* dos modelos com distribuições de probabilidade testadas.

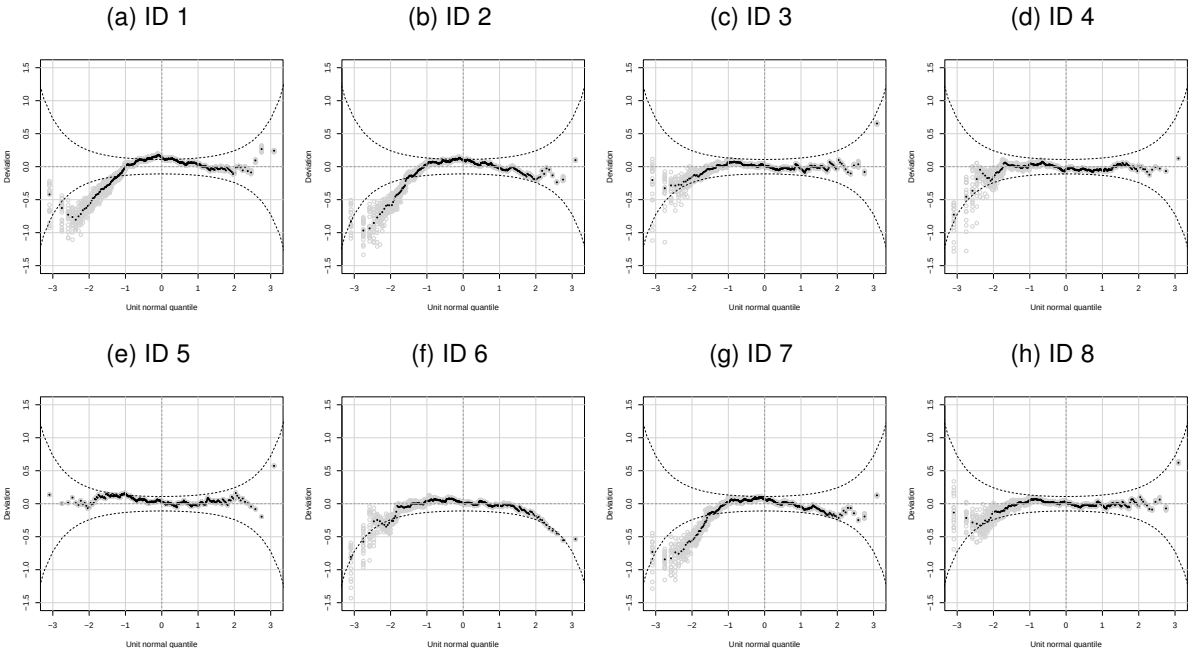
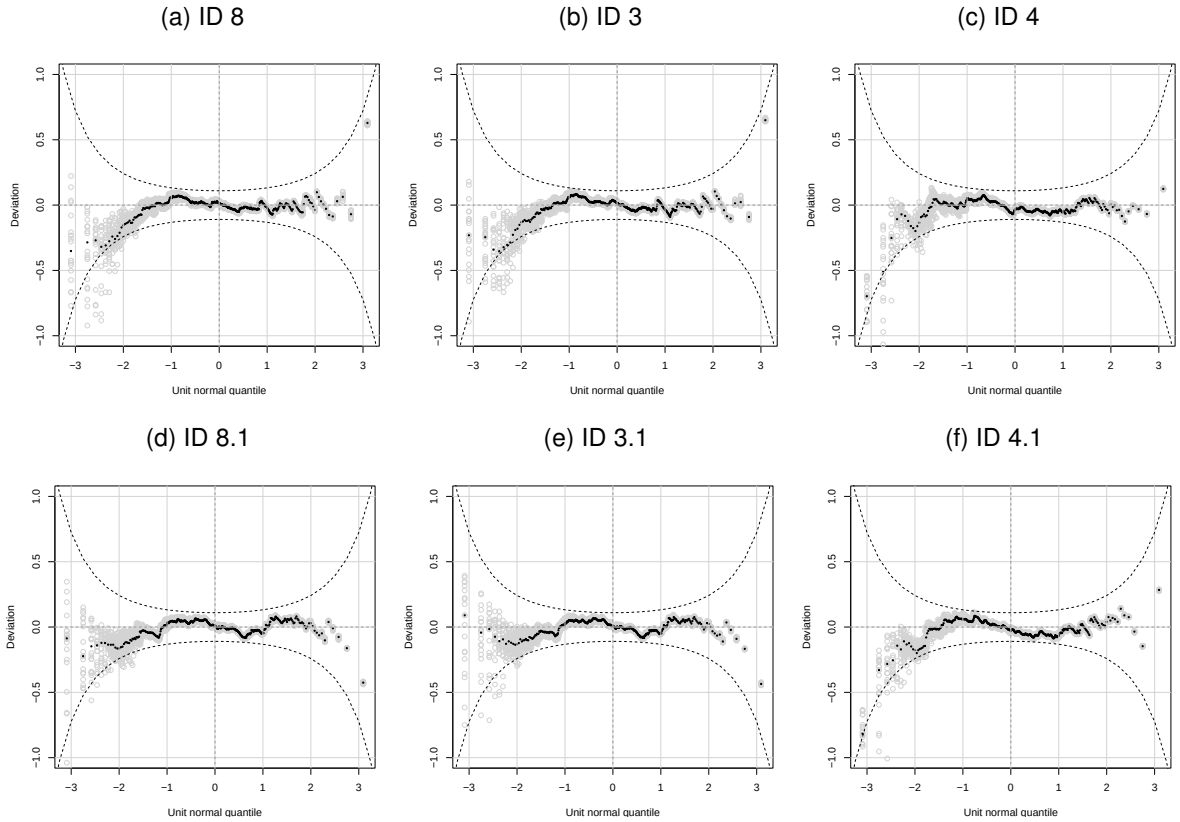


Tabela 3.6 – Modelos ajustados com distribuições Delaporte, Binomial Negativa Tipo I e II, com preditor linear para μ.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
8	Delaporte	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4458,47
8.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4565,43
3	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4459,08
3.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4563,74
4	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4462,06
4.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4474,11

É possível verificar tanto pelo *worm plot* quanto pelo teste da razão de verossimilhança que os modelos 8, 4 e 3, com ponto e altura como fatores de efeitos aleatórios, são mais apropriados. Assim, considerando tal estrutura aleatória, o parâmetro σ foi modelado via preditor linear em uma tentativa de melhorar o ajuste do modelo, uma vez que na análise descritiva foi verificado que a variabilidade parece ser diferente com relação aos níveis dos fatores. Além disso, foi possível identificar duas observações que podem ser consideradas outliers nos modelos, observações 164 e 500. Como a remoção de observações de um banco de dados deve ser uma medida feita com cautela, especialmente quando o banco de dados é considerado pequeno, foram testadas a remoção dos outliers de forma separada e removendo apenas um deles em cada vez. Com isso, verificou-se que com

Figura 3.11 – *Worm plot* dos modelos com distribuições de probabilidade pré-selecionadas, com preditor linear para μ e diferentes estruturas aleatórias.



a remoção da observação 500, o modelo foi capaz de se ajustar adequadamente sem grandes problemas. Desta forma, considera-se os modelos com

$$\begin{aligned} g_1(\mu_{ijkt}) = \eta^{\mu}_{ijkt} &= \beta_0 + \alpha_i + \theta_j + \gamma_t + b_{(j|k)} + b_k \\ &= (\beta_0 + b_{(j|k)} + b_k) + \alpha_i + \theta_j + \gamma_t \end{aligned} \tag{3.2.2}$$

$$g_2(\sigma_{ijkt}) = \eta^{\sigma}_{ijkt} = \beta_0 + \alpha_i + \theta_j + \gamma_t.$$

Tabela 3.7 – Teste da razão de verossimilhança entre os modelos com diferentes estruturas aleatórias.

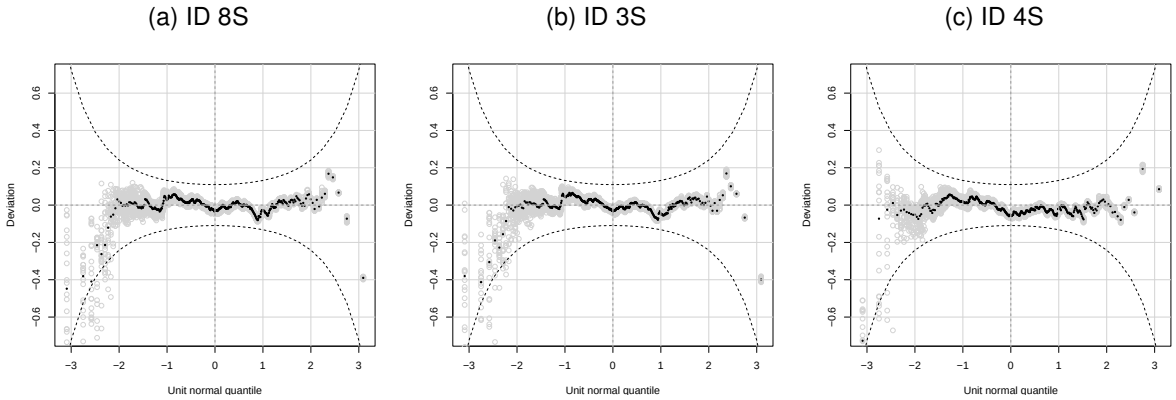
Modelo 8.1 (Nulo):	Deviance = 4519,28	com 23,077 graus de liberdade
Modelo 8 (Alternativo):	Deviance = 4381,06	com 38,7043 graus de liberdade
LRT = 138,2165 com 15,6273 graus de liberdade		
p-valor < 0,0001		
Modelo 3.1 (Nulo):	Deviance = 4519,25	com 22,2473 graus de liberdade
Modelo 3 (Alternativo):	Deviance = 4383,74	com 37,6719 graus de liberdade
LRT = 135,5097 com 15,4245 graus de liberdade		
p-valor < 0,0001		
Modelo 4.1 (Nulo):	Deviance = 4423,53	com 25,2906 graus de liberdade
Modelo 4 (Alternativo):	Deviance = 4389,15	com 36,4548 graus de liberdade
LRT = 34,3784 com 11,1642 graus de liberdade		
p-valor < 0,001		

Tabela 3.8 – Modelos ajustados com distribuições Delaporte, Binomial Negativa Tipo I e II, com e sem preditor linear para σ.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
8	Delaporte	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4432,59
8S		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	4413,47
3	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4433,18
3S		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	4413,26
4	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	4445,73
4S		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	4430,52

Comparando o AIC dos modelos com e sem preditor linear para σ (Tabela 3.8), tem-se que a inclusão da modelagem de σ pelos fatores do experimento contribui para um ajuste melhor, corroborado pelos gráficos *worm plots* (Figura 3.12) que apresentaram evolução, corrigindo uma curvatura para baixo na cauda esquerda que se apresentava anteriormente. Vale destacar que o modelo sem o preditor linear para σ, mostrado na Tabela 3.8, foi reajustado novamente após a remoção do outlier, permitindo a comparação entre os modelos pelo critério AIC a partir do mesmo banco de dados. Nesse sentido, apesar de os gráficos *worm plots* serem muito similares, pode-se inferir que o ajuste pela Binomial Negativa Tipo II teve um desempenho inferior em comparação com as outras duas, tanto na modelagem apenas da média quanto na modelagem do desvio padrão também, tendo o maior AIC

Figura 3.12 – *Worm plot* dos modelos com distribuições de probabilidade pré-selecionadas, com preditor linear para μ e σ.

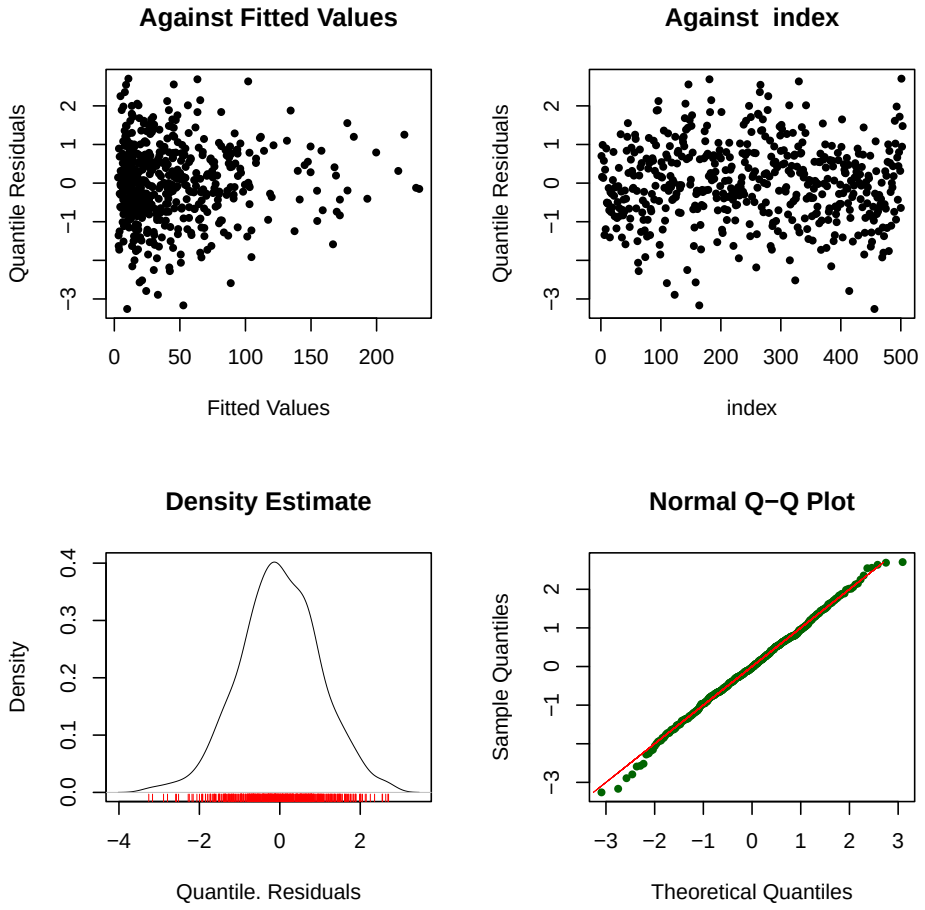


de todos. Além disso, a distribuição Delaporte apresentou resultados muito similares aos da Binomial Negativa Tipo I, em todos os aspectos, sendo que a estimativa do parâmetro populacional ν foi de $2,22045 \cdot 10^{-16}$, isto é, $\nu \approx 0$. Portanto, a relação entre a média e variância da resposta se mostrou mais adequada na forma quadrática $Var(Y) = \mu + \sigma\mu^2$, e com isso, dado que o ajuste pode ser considerado como satisfatório, a distribuição de probabilidade Binomial Negativa Tipo I é a melhor opção dentre as distribuições testadas.

Tabela 3.9 – Medidas descritivas dos resíduos simulados do modelo Binomial Negativa Tipo I.

Média	> -0,005
Variância	1,011
Coeficiente de Assimetria	-0,087
Coeficiente de Curtose	3,192

Figura 3.13 – Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo I.



Na Figura 3.13, verifica-se a conformidade dos resíduos aleatorizados, que se comportam de forma aleatória em torno de zero. Pelas medidas descritivas, vide Tabela 3.9, a média, variância, assimetria e curtose estão próximas de 0, 1, 0 e 3, respectivamente, indicando uma boa aproximação de uma distribuição normal padrão.

As estimativas dos coeficientes de regressão, apresentadas na Tabela 3.10, indicam que para o preditor de μ, todas as estimativas dos coeficientes de talhão são negativas, isto é, a média da contagem de insetos é menor nos outros talhões em relação ao *baseline*,

Tabela 3.10 – Estimativas dos coeficientes de regressão e parâmetros de variância do modelo Binomial Negativa Tipo I.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($> \|t\|$)
			μ		
(Intercepto)	$\boldsymbol{\beta_0}$	**3,62**	**0,11**	**31,84**	**< 0,001**
Talhão 2	α_2	-0,06	0,09	-0,61	0,54
Talhão 3	$\boldsymbol{\alpha_3}$	**-0,86**	**0,12**	**-7,16**	**< 0,001**
Talhão 4	α_4	-0,14	0,10	-1,38	0,17
Talhão 5	$\boldsymbol{\alpha_5}$	**-0,97**	**0,13**	**-7,63**	**< 0,001**
Talhão 6	$\boldsymbol{\alpha_6}$	**-1,87**	**0,11**	**-16,31**	**< 0,001**
Altura 4,5	θ_2	0,08	0,08	1,03	0,30
Altura 7	θ_3	0,03	0,09	0,38	0,70
Coleta 2	$\boldsymbol{\gamma_2}$	**1,19**	**0,12**	**9,72**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,69**	**0,11**	**6,05**	**< 0,001**
Coleta 4	$\boldsymbol{\gamma_4}$	**0,53**	**0,13**	**4,12**	**< 0,001**
Coleta 5	γ_5	-0,15	0,14	-1,02	0,31
Coleta 6	γ_6	0,19	0,13	1,43	0,15
Coleta 7	$\boldsymbol{\gamma_7}$	**0,38**	**0,13**	**2,86**	**< 0,005**
			σ		
(Intercepto)	$\boldsymbol{\beta_0}$	**-1,32**	**0,28**	**-4,66**	**< 0,001**
Talhão 2	α_2	0,52	0,27	1,90	0,06
Talhão 3	$\boldsymbol{\alpha_3}$	**1,32**	**0,25**	**5,24**	**< 0,001**
Talhão 4	$\boldsymbol{\alpha_4}$	**0,84**	**0,26**	**3,25**	**< 0,005**
Talhão 5	$\boldsymbol{\alpha_5}$	**1,41**	**0,26**	**5,52**	**< 0,001**
Talhão 6	$\boldsymbol{\alpha_6}$	**1,02**	**0,28**	**3,58**	**< 0,001**
Altura 4,5	θ_2	-0,25	0,17	-1,45	0,15
Altura 7	θ_3	0,05	0,17	0,29	0,77
Coleta 2	γ_2	-0,15	0,25	-0,59	0,56
Coleta 3	$\boldsymbol{\gamma_3}$	**-0,70**	**0,26**	**-2,65**	**< 0,01**
Coleta 4	γ_4	-0,07	0,26	-0,25	0,80
Coleta 5	γ_5	0,30	0,27	1,13	0,26
Coleta 6	γ_6	-0,06	0,26	-0,22	0,83
Coleta 7	γ_7	-0,11	0,26	-0,41	0,68

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,2615	0,5114
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,0162	0,1274
Resíduo	ϵ	1,0016	1,0008

no entanto, os talhões 2 e 4 não foram significativos. Nenhum dos níveis de altura foram significativos e em relação a coleta, apenas as coletas 5 e 6 não foram significativas. Já com relação ao preditor de σ, tanto os níveis de altura quanto os níveis de coleta não foram

significativos perante o *baseline*, com exceção da coleta 3, e com relação ao fator talhão, apenas o talhão 2 não foi significativo mas com um p-valor baixo, de 0,06. Desta forma, optou-se por testar a remoção dos fatores não significativos em busca de uma melhora e simplificação do modelo, conforme indica a Tabela 3.11.

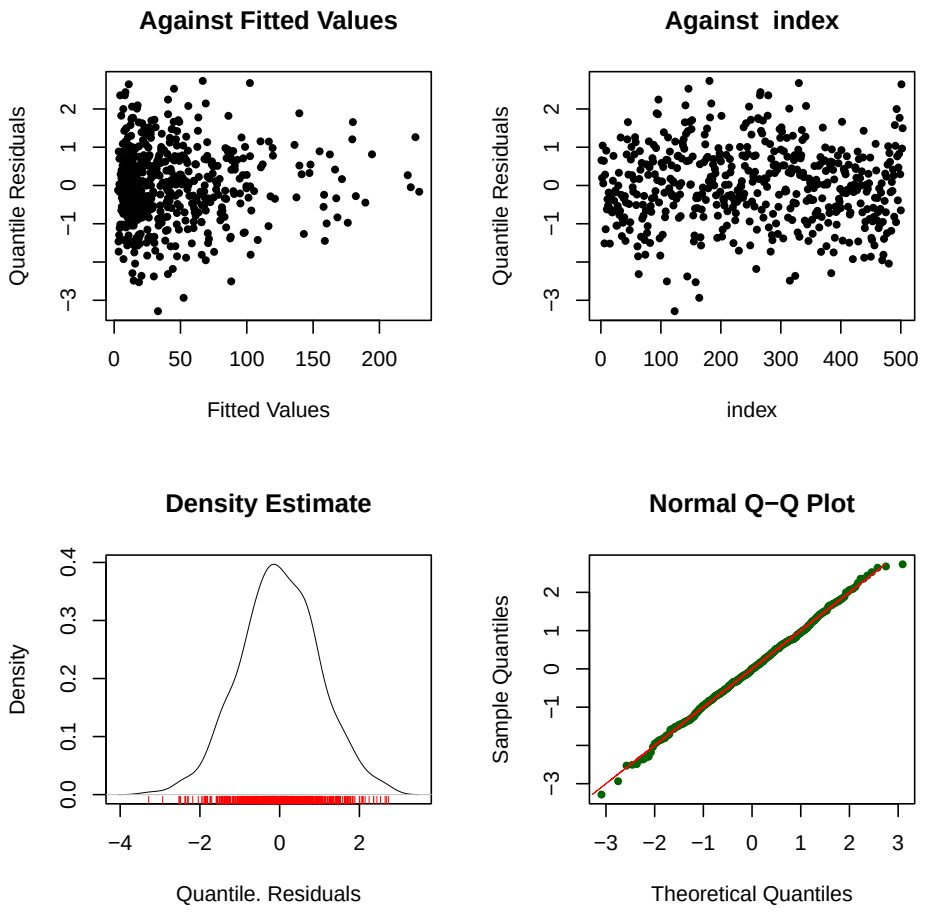
Tabela 3.11 – Modelos testados com a distribuição Binomial Negativa Tipo I.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
3S	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	4413,26
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j$	4413,29
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \gamma_t$	4411,67
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	4410,04
		$\boldsymbol{(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t}$	$\boldsymbol{\beta_0 + \alpha_i + \gamma_t}$	**4408,39**
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \theta_j$	4409,98
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i$	4407,89

Seguindo a mesma metodologia de seleção de modelos para a comparação de diferentes distribuições de probabilidade, os modelos testados (Tabela 3.11) com a distribuição Binomial Negativa Tipo I, a partir da remoção dos fatores não significativos, foram comparados a partir do critério AIC e da inspeção visual dos *worm plots*. A partir disso, o modelo que melhor se ajustou aos dados, escolhido como modelo final, contém os fatores de efeitos fixos talhão e coleta para os preditores de μ e σ, com ponto e altura como fatores de efeitos aleatórios para μ. De acordo com os coeficientes de regressão apresentados na Tabela 3.12, as estimativas e conclusões dos testes seguem em linha com o modelo inicial (Tabela 3.10), sendo que o talhão 2, para o preditor de σ, passou a ser significativo.

Os gráficos de diagnóstico e *worm plot* (Figuras 3.14-3.15) indicam que o ajuste é adequado, aparentando evolução quando comparado com o modelo inicial. As medidas descritivas dos resíduos simulados (Tabela 3.13) estão próximas de uma normal padrão e, neste caso, o coeficiente de curtose apresentou melhora. Conclui-se, portanto, que o modelo proposto é adequado para o ajuste dos dados do estudo e pode ser escritos como:

Figura 3.14 – Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo I - modelo final.



$$
\begin{aligned}
log(\hat{\mu}_{ijkt}) = \hat{\eta}^{\mu}_{ijkt} &= (3,65 + \hat{b}_k + \hat{b}_{(j|k)}) - 0,05\, Talhão\ 2 - 0,86\, Talhão\ 3 + \\
&- \ 0,13\, Talhão\ 4 - 0,98\, Talhão\ 5 - 1,88\, Talhão\ 6 + \\
&+ \ 1,17\, Coleta\ 2 + 0,69\, Coleta\ 3 + 0,53\, Coleta\ 4 + \\
&- \ 0,14\, Coleta\ 5 + 0,20\, Coleta\ 6 + 0,37\, Coleta\ 7 \\
log(\hat{\sigma}_{ijkt}) = \hat{\eta}^{\sigma}_{ijkt} &= -1,50 + 0,59\, Talhão\ 2 + 1,41\, Talhão\ 3 + 1,01\, Talhão\ 4 + \\
&+ \ 1,55\, Talhão\ 5 + 1,16\, Talhão\ 6 - 0,13\, Coleta\ 2 + \\
&- \ 0,69\, Coleta\ 3 - 0,04\, Coleta\ 4 + 0,28\, Coleta\ 5 + \\
&- \ 0,05\, Coleta\ 6 - 0,10\, Coleta\ 7.
\end{aligned}
\tag{3.2.3}
$$

A partir do modelo proposto e das equações apresentadas em (3.2.3), pode-se concluir que os talhões possuem efeito negativo no logaritmo da média em relação ao seu *baseline*, enquanto para as coletas, o efeito é positivo e é possível observar a ciclicidade pela dimen-

Tabela 3.12 – Estimativas dos coeficientes de regressão e parâmetros de variância do modelo final - Binomial Negativa Tipo I.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($> \|t\|$)
			μ		
(Intercepto)	$\boldsymbol{\beta_0}$	**3,65**	**0,10**	**36,13**	**< 0,001**
Talhão 2	α_2	-0,05	0,09	-0,54	0,59
Talhão 3	$\boldsymbol{\alpha_3}$	**-0,86**	**0,12**	**-7,35**	**< 0,001**
Talhão 4	α_4	-0,13	0,10	-1,28	0,20
Talhão 5	$\boldsymbol{\alpha_5}$	**-0,98**	**0,13**	**-7,69**	**< 0,001**
Talhão 6	$\boldsymbol{\alpha_6}$	**-1,88**	**0,11**	**-16,57**	**< 0,001**
Coleta 2	$\boldsymbol{\gamma_2}$	**1,17**	**0,12**	**9,76**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,69**	**0,11**	**6,10**	**< 0,001**
Coleta 4	$\boldsymbol{\gamma_4}$	**0,53**	**0,13**	**4,24**	**< 0,001**
Coleta 5	γ_5	-0,14	0,14	-0,98	0,33
Coleta 6	γ_6	0,20	0,13	1,58	0,12
Coleta 7	$\boldsymbol{\gamma_7}$	**0,37**	**0,13**	**2,90**	**< 0,005**
			σ		
(Intercepto)	$\boldsymbol{\beta_0}$	**-1,50**	**0,26**	**-5,74**	**< 0,001**
Talhão 2	$\boldsymbol{\alpha_2}$	**0,59**	**0,27**	**2,21**	**0,03**
Talhão 3	$\boldsymbol{\alpha_3}$	**1,41**	**0,25**	**5,74**	**< 0,001**
Talhão 4	$\boldsymbol{\alpha_4}$	**1,01**	**0,25**	**4,11**	**< 0,001**
Talhão 5	$\boldsymbol{\alpha_5}$	**1,55**	**0,25**	**6,22**	**< 0,001**
Talhão 6	$\boldsymbol{\alpha_6}$	**1,16**	**0,27**	**4,24**	**< 0,001**
Coleta 2	γ_2	-0,13	0,25	-0,53	0,60
Coleta 3	$\boldsymbol{\gamma_3}$	**-0,69**	**0,26**	**-2,62**	**< 0,01**
Coleta 4	γ_4	-0,04	0,26	-0,16	0,87
Coleta 5	γ_5	0,28	0,27	1,05	0,29
Coleta 6	γ_6	-0,05	0,26	-0,18	0,86
Coleta 7	γ_7	-0,10	0,26	-0,39	0,70

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,2723	0,5218
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,0182	0,1348
Resíduo	ϵ	1,0072	1,0036

são dos coeficientes, que decresce da coleta 2 até a coleta 5 e volta a crescer novamente, sendo que o logaritmo da média na coleta 2 é menor do que nos demais. Considerando os efeitos aleatórios associados a ponto, no nível de indivíduo, b_k, e associado ao fator altura dentro de ponto, $b_{(j|k)}$, os desvios padrão foram de 0,5218 e 0,1348, respectivamente, o que representa cerca de 39,5% da variabilidade total atribuída a esses efeitos.

Em relação ao logaritmo do desvio, a partir do modelo, tem-se que a variabilidade é

Figura 3.15 – *Worm plot* do modelo final - Binomial Negativa Tipo I.

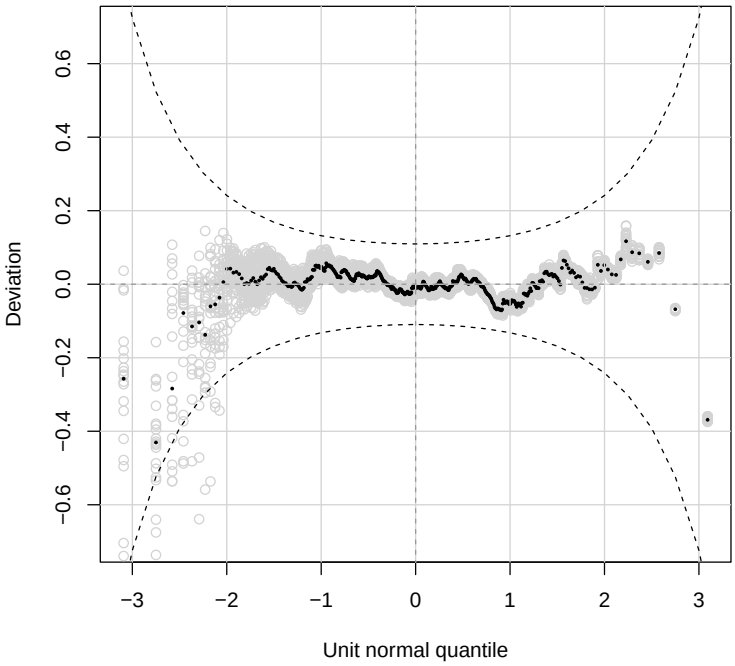


Tabela 3.13 – Medidas descritivas dos resíduos simulados do modelo final - Binomial Negativa Tipo I.

Média	-0,003
Variância	0,996
Coeficiente de Assimetria	-0,025
Coeficiente de Curtose	3,07

maior nos demais talhões em relação ao baseline, com variabilidade maior no talhão 5. Para o fator coleta, o efeito no logaritmo do desvio é negativo, com exceção da coleta 5. Isso quer dizer que a variabilidade nos demais talhões, com o passar das coletas, é menor em relação ao baseline exceto pela coleta 5, que tem variabilidade maior. Assim, espera-se que a maior variabilidade seja encontrada no talhão 5 (localização 2) e na coleta 5.

Conforme apresentado na Tabela 3.14, cada talhão foi colocado como referência e comparado com os demais em relação à significância da diferença entre eles. Com isso, é possível observar que não há diferença significativa apenas entre os talhões 1 e 2; 1 e 4; 2 e 4; e 3 e 5, lembrando que os talhões 1, 2 e 3 pertencem à localização 1 e os demais, à localização 2. Portanto, na localização 1, o talhão 3 possui contagem média significativamente menor que os outros dois, que não se diferem entre si. Já na localização 2, todos os talhões se diferem entre si, não havendo diferenças apenas com relação aos

Tabela 3.14 – Estimativas dos coeficientes de talhão com diferentes *baselines*.

Efeitos Fixos		Estimativa (p-valor)								
		Baseline 1			*Baseline* 2			*Baseline* 3		
Talhão 1	α_1	(Intercepto)			0,05	(0,59)		0,86	(< 0,001)	**(*)**
Talhão 2	α_2	-0,05	(0,59)		(Intercepto)			0,81	(< 0,001)	**(*)**
Talhão 3	α_3	-0,86	(< 0,001)	**(*)**	-0,81	(< 0,001)	**(*)**	(Intercepto)		
Talhão 4	α_4	-0,13	(0,20)		-0,08	(0,46)		0,73	(< 0,001)	**(*)**
Talhão 5	α_5	-0,98	(< 0,001)	**(*)**	-0,93	(< 0,001)	**(*)**	-0,11	(0,47)	
Talhão 6	α_6	-1,88	(< 0,001)	**(*)**	-1,83	(< 0,001)	**(*)**	-1,01	(< 0,001)	**(*)**
		Baseline 4			*Baseline* 5			*Baseline* 6		
Talhão 1	α_1	0,13	(0,20)		0,98	(< 0,001)	**(*)**	1,88	(< 0,001)	**(*)**
Talhão 2	α_2	0,08	(0,46)		0,93	(< 0,001)	**(*)**	1,83	(< 0,001)	**(*)**
Talhão 3	α_3	-0,73	(< 0,001)	**(*)**	0,11	(0,47)		1,01	(< 0,001)	**(*)**
Talhão 4	α_4	(Intercepto)			0,85	(< 0,001)	**(*)**	1,75	(< 0,001)	**(*)**
Talhão 5	α_5	-0,85	(< 0,001)	**(*)**	(Intercepto)			0,90	(< 0,001)	**(*)**
Talhão 6	α_6	-1,75	(< 0,001)	**(*)**	-0,90	(< 0,001)	**(*)**	(Intercepto)		

(*) – Coeficientes significativos em relação ao *baseline*.

talhões da localização 1, uma vez que os talhões 1 e 4; 2 e 4; e 3 e 5 não se diferem significativamente. Ainda, pode-se constatar que, com exceção do talhão 4 em relação ao talhão 3, todos os coeficientes associados a localização 2 são negativos. Portanto, embora a contagem média no talhão 4 seja maior do que no talhão 3, há evidências de que a resposta média na localização 2 seja diferente, a menor, da localização 1. Desta forma, partindo do modelo final, compara-se os talhões entre as duas localizações, agrupando-os por meio de constrastes e testando a seguinte hipótese:

$$\begin{aligned} H_0 &: (\alpha_1 + \alpha_2 + \alpha_3) - (\alpha_4 + \alpha_5 + \alpha_6) = 0, \quad \text{não há diferença entre localizações} \\ H_1 &: (\alpha_1 + \alpha_2 + \alpha_3) - (\alpha_4 + \alpha_5 + \alpha_6) \neq 0, \quad \text{há diferença entre localizações.} \end{aligned} \tag{3.2.4}$$

Com base no agrupamento dos talhões por contrastes e o teste de hipótese apresentado na Equação (3.2.4), apresenta-se, na Tabela 3.15, as estimativas dos coeficientes do modelo ajustado. Conforme presumido, o efeito associado à localização 2 é negativo e o teste de hipótese aponta diferença significativa entre os grupos, tanto na contagem média quando na variabilidade. Portanto, considerando que a localização é significativa na modelagem conjunta e a distribuição da contagem de insetos difere de acordo com o pomar, as localizações serão modeladas separadamente com o propósito de melhorar o ajuste do modelo e investigar qual distribuição de probabilidade melhor se ajusta em cada caso.

Tabela 3.15 – Estimativas dos coeficientes de regressão e parâmetros de variância do modelo final com talhões agrupados - Binomial Negativa Tipo I.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($> \|t\|$)
		μ			
(Intercepto)	$\boldsymbol{\beta_0}$	**3,04**	**0,09**	**32,86**	**< 0,001**
Localização 2 (T4, T5 e T6)	$\boldsymbol{\alpha_2}$	**-0,61**	**0,13**	**-4,65**	**< 0,001**
Coleta 2	$\boldsymbol{\gamma_2}$	**1,12**	**0,13**	**8,72**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,67**	**0,12**	**5,51**	**< 0,001**
Coleta 4	$\boldsymbol{\gamma_4}$	**0,65**	**0,14**	**4,48**	**< 0,001**
Coleta 5	γ_5	-0,13	0,15	-0,89	0,37
Coleta 6	γ_6	0,17	0,14	1,16	0,25
Coleta 7	$\boldsymbol{\gamma_7}$	**0,50**	**0,14**	**3,61**	**< 0,001**
		σ			
(Intercepto)	$\boldsymbol{\beta_0}$	**-0,57**	**0,18**	**-3,11**	**< 0,005**
Localização 2 (T4, T5 e T6)	$\boldsymbol{\alpha_2}$	**-0,77**	**0,32**	**-2,40**	**0,02**
Coleta 2	γ_2	0,02	0,25	0,07	0,94
Coleta 3	γ_3	-0,36	0,25	-1,41	0,16
Coleta 4	γ_4	0,43	0,25	1,75	0,08
Coleta 5	γ_5	0,40	0,25	1,58	0,11
Coleta 6	γ_6	0,39	0,25	1,56	0,12
Coleta 7	γ_7	0,24	0,25	0,98	0,33

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,6143	0,7838
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,0275	0,1658
Resíduo	ϵ	0,9920	0,9960

3.3 Modelagem 2 – Localização 1

Considerando a localização 1, foram selecionadas apenas as distribuições Binomial Negativa Tipo I e Tipo II para a modelagem dos dados, visto que tais distribuições tiveram melhor aderência na modelagem conjunta. Inicialmente, foram considerados modelos lineares generalizados mistos modelando apenas a média via preditor linear e testados ponto e ponto e altura como fatores de efeitos aleatórios, conforme Tabela 3.16.

Os gráficos *worm plots* dos modelos (Figura 3.16) sinalizam que o ajuste não foi adequado, entretanto, há um outlier (observação 164) prejudicando a análise diagnóstica. Desta forma, o outlier foi removido e o preditor linear para σ foi incluido para o reajuste do modelo.

Figura 3.16 – *Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ.

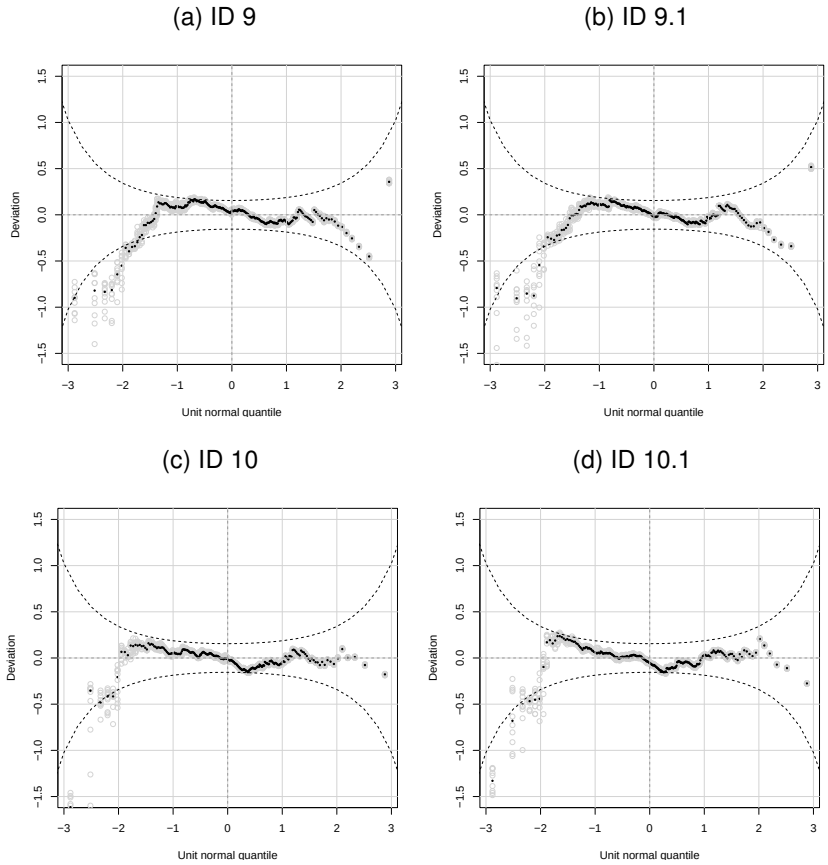


Tabela 3.16 – Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
9	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2342,77
9.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2348,15
10	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2306,11
10.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2327,20

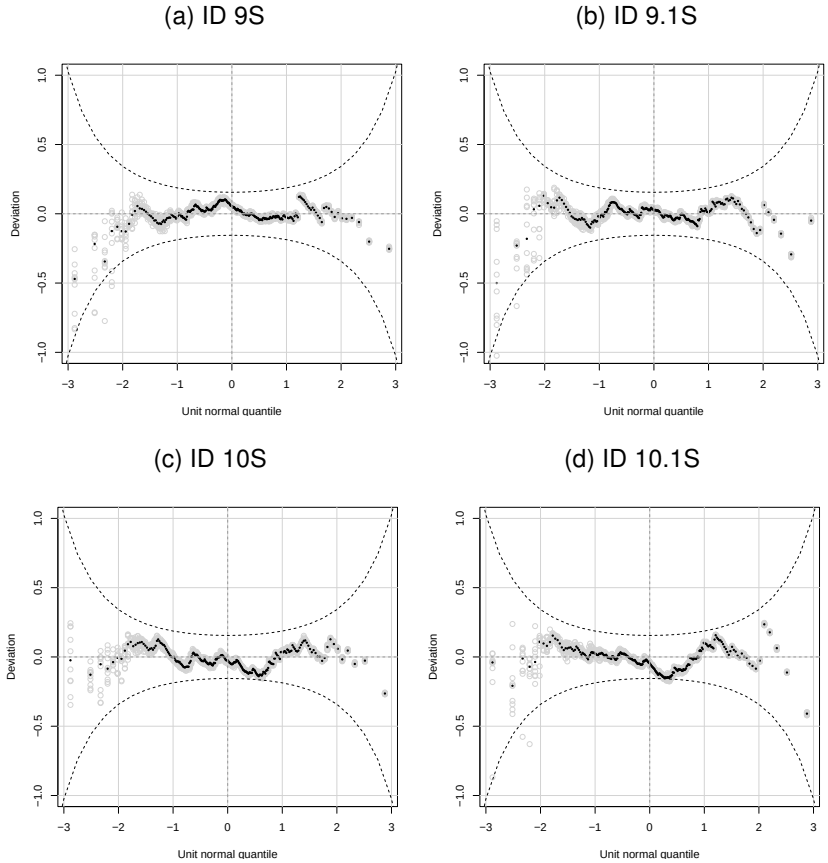
De acordo com a Figura 3.17, nota-se que a remoção do outlier e a inclusão do preditor linear para σ foi suficiente para melhorar consideravelmente o ajuste dos modelos, sendo que todos eles aparentam um ajuste adequado. Pelo critério AIC, presente na Tabela 3.17, os modelos contendo ponto e altura como efeitos aleatórios são considerados melhores e, para verificar se a diferença entre os modelos é significativa, utiliza-se o teste da razão de verossimilhança.

Pelo teste da razão de verossimilhança entre os modelos com diferentes estruturas

Tabela 3.17 – Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
9S	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	2302,46
9.1S		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	2309,96
10S	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	2276,79
10.1S		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	2304,74

Figura 3.17 – *Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ.



aleatórias, apresentado na Tabela 3.18, conclui-se que os modelos diferem significativamente e a estrutura com ponto e altura como fatores de efeitos aleatórios é preferida, isto é, modelos 1 e 2. Desta forma, é necessária a comparação direta entre os dois melhores modelos para verificar qual das relações entre a média e variância é mais adequada para a ocasião.

Na conjuntura da análise diagnóstica, tem-se que o modelo 2 fornece um ajuste melhor corroborado pelo *worm plot* (Figura 3.18), que não apresenta nenhum padrão e o menor

Tabela 3.18 – Teste da razão de verossimilhança entre os modelos com diferentes estruturas aleatórias.

Modelo 9.1S (Nulo):	Deviance = 2245,013	com 32,4758 graus de liberdade
Modelo 9S (Alternativo):	Deviance = 2217,144	com 42,6593 graus de liberdade
LRT = 27,8685 com 10,1835 graus de liberdade		
p-valor = 0,0021165		

Modelo 10.1S (Nulo):	Deviance = 2240,227	com 32,2573 graus de liberdade
Modelo 10S (Alternativo):	Deviance = 2187,701	com 44,5442 graus de liberdade
LRT = 52,6266 com 12,2869 graus de liberdade		
p-valor < 0,0001		

Figura 3.18 – Modelos com distribuição de probabilidade Binomial Negativa Tipo I e Tipo II com melhor desempenho.

(a) ID 9S (b) ID 10S

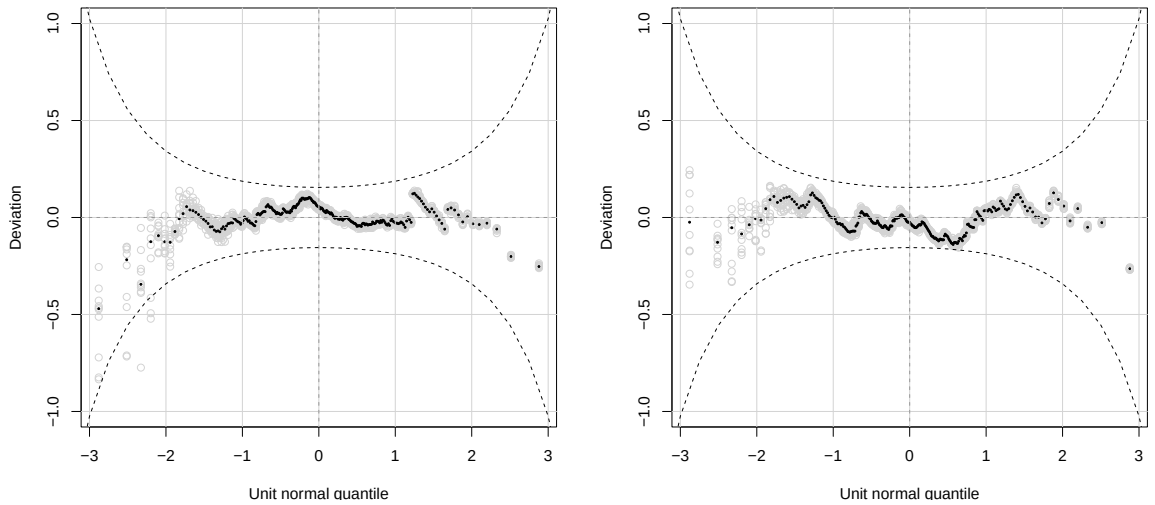
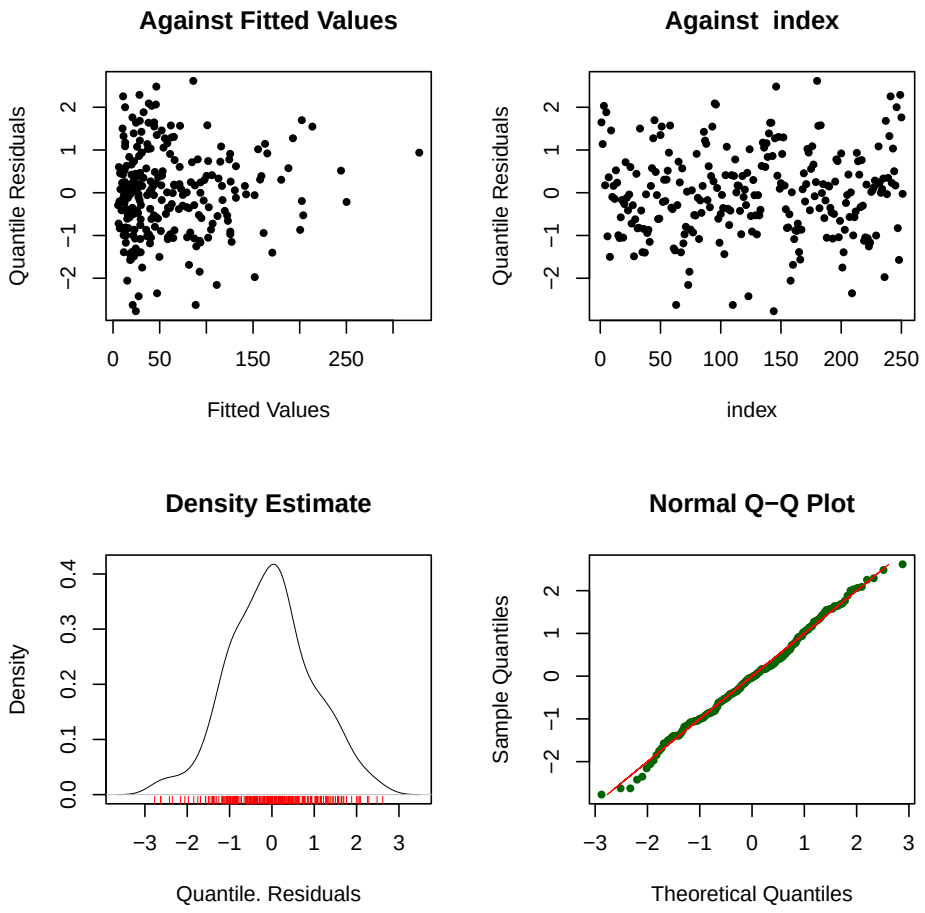


Tabela 3.19 – Medidas descritivas dos resíduos simulados do modelo Binomial Negativa Tipo II.

Média	-0,023
Variância	1,005
Coeficiente de Assimetria	0,042
Coeficiente de Curtose	2,98

AIC dentre as distribuições testadas. O *plot* do modelo (Figura 3.19) também apresenta conformidade com os pressupostos e os resíduos aleatorizados estão próximos de uma distribuição normal padrão (Tabela 3.19), como esperado. Com relação aos coeficientes de regressão do modelo, para a média, apenas o fator altura não difere significativamente

Figura 3.19 – Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II.



entre os níveis, enquanto que para o desvio, o mesmo ocorre tanto para altura quanto para talhão.

Até aqui, em comparação ao conjunto de dados completo, pode-se concluir que a relação média-variância na forma $Var(Y) = \mu + \sigma\mu$ se adequa melhor ao conjunto de dados referente a localização 1 do que na forma quadrática. Além disso, a significância dos níveis dos fatores também se diferem em cada caso, principalmente com relação à predição de σ, uma vez que, na modelagem conjunta, os níveis de talhão foram altamente significativos e os níveis de coleta, em sua maioria, foram não significativos. Já para a localização 1, ocorre justamente o contrário: os níveis de talhão não foram significativos e os níveis de coleta, com exceção da coleta 3, foram todos significativos.

Tabela 3.20 – Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($> \|t\|$)
			μ		
(Intercepto)	$\boldsymbol{\beta_0}$	**3,31**	**0,09**	**37,08**	**< 0,001**
Talhão 2	α_2	-0,09	0,08	-1,17	0,24
Talhão 3	$\boldsymbol{\alpha_3}$	**-1,01**	**0,10**	**-9,93**	**< 0,001**
Altura 4,5	θ_2	-0,04	0,07	-0,51	0,61
Altura 7	θ_3	0,00	0,07	-0,02	0,98
Coleta 2	$\boldsymbol{\gamma_2}$	**1,36**	**0,11**	**12,85**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,93**	**0,10**	**9,16**	**< 0,001**
Coleta 4	$\boldsymbol{\gamma_4}$	**0,88**	**0,12**	**7,54**	**< 0,001**
Coleta 5	γ_5	0,23	0,13	1,75	0,08
Coleta 6	$\boldsymbol{\gamma_6}$	**0,67**	**0,11**	**6,22**	**< 0,001**
Coleta 7	$\boldsymbol{\gamma_7}$	**0,59**	**0,13**	**4,50**	**< 0,001**
			σ		
(Intercepto)	$\boldsymbol{\beta_0}$	**1,23**	**0,36**	**3,39**	**< 0,001**
Talhão 2	α_2	-0,05	0,33	-0,15	0,88
Talhão 3	α_3	0,13	0,31	0,41	0,68
Altura 4,5	θ_2	0,26	0,25	1,02	0,31
Altura 7	θ_3	0,30	0,27	1,09	0,28
Coleta 2	$\boldsymbol{\gamma_2}$	**1,56**	**0,41**	**3,77**	**< 0,001**
Coleta 3	γ_3	0,53	0,47	1,13	0,26
Coleta 4	$\boldsymbol{\gamma_4}$	**1,34**	**0,45**	**3,00**	**< 0,005**
Coleta 5	$\boldsymbol{\gamma_5}$	**1,09**	**0,42**	**2,62**	**0,01**
Coleta 6	$\boldsymbol{\gamma_6}$	**1,04**	**0,41**	**2,55**	**0,01**
Coleta 7	$\boldsymbol{\gamma_7}$	**1,18**	**0,47**	**2,52**	**0,01**

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,3840	0,6197
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,0570	0,2388
Resíduo	ϵ	1,1035	1,0505

A partir do modelo com distribuição de probabilidade Binomial Negativa Tipo II, ponto e altura como fatores de efeitos aleatórios e preditor linear para μ e σ, busca-se por um modelo mais parcimonioso a partir da remoção de alguns fatores, em particular, aqueles que não apresentaram diferenças significativas. Na Tabela 3.21, observa-se que os modelos apresentam critérios AIC muito próximos entre si, com uma amplitude de 10 unidades entre o maior e o menor critério. Com base nisso, a escolha pelo modelo destacado, removendo o fator altura em ambos os preditores, se deu pelo comportamento dos resíduos

Tabela 3.21 – Modelos testados com a distribuição Binomial Negativa Tipo II.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	2276,79
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	2272,69
10S	Binomial Negativa Tipo II	$\boldsymbol{(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t}$	$\boldsymbol{\beta_0 + \alpha_i + \gamma_t}$	**2270,50**
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \theta_j + \gamma_t$	2268,74
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \gamma_t$	2266,64

em uma inspeção gráfica (Figuras 3.20-3.21) e suas medidas descritivas (Tabela 3.22). De forma similar ao que ocorre na modelagem conjunta, a presença do fator talhão no modelo para o preditor de σ é necessária para a conformidade do modelo, mesmo que os níveis não difiram entre si. Portanto, tem-se as funções de ligação:

$$\begin{aligned} log(\hat{\mu}_{ijkt}) = \hat{\eta}^{\mu}_{ijkt} &= (3,36 + \hat{b}_k + \hat{b}_{(j|k)}) - 0,07\, Talhão\ 2 - 0,99\, Talhão\ 3 + \\ &+\ 1,36\, Coleta\ 2 + 0,92\, Coleta\ 3 + 0,85\, Coleta\ 4 + \\ &+\ 0,23\, Coleta\ 5 + 0,65\, Coleta\ 6 + 0,55\, Coleta\ 7 \\ log(\hat{\sigma}_{ijkt}) = \hat{\eta}^{\sigma}_{ijkt} &= 1,45 - 0,03\, Talhão\ 2 + 0,17\, Talhão\ 3 + 1,51\, Coleta\ 2 + \\ &+\ 0,51\, Coleta\ 3 + 1,30\, Coleta\ 4 + 1,05\, Coleta\ 5 + \\ &+\ 0,97\, Coleta\ 6 + 1,07\, Coleta\ 7. \end{aligned} \tag{3.3.1}$$

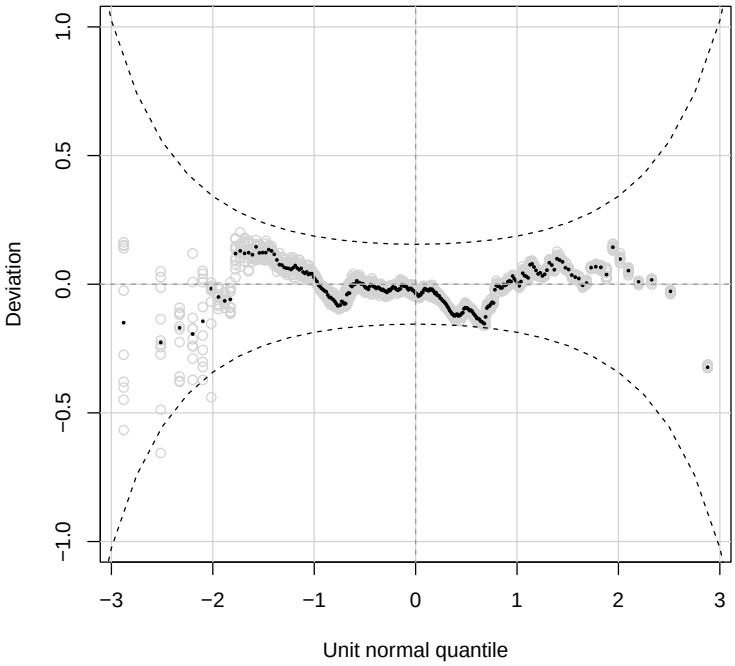
Tabela 3.22 – Medidas descritivas dos resíduos simulados do modelo final - Binomial Negativa Tipo II.

Média	-0,022
Variância	1,004
Coeficiente de Assimetria	0,025
Coeficiente de Curtose	3,03

Interpretando os coeficientes de regressão do modelo, a partir da Equação (3.3.1), em relação ao *baseline* talhão 1 e coleta 1, observam-se efeitos negativos nos demais talhões para o preditor de μ, sendo que a menor média ocorre no talhão 3, como observado anteriormente também na análise exploratória dos dados. As estimativas dos coeficientes de regressão associados ao fator coleta são todos positivos, indicando que a contagem média é menor na coleta 1 e maior na coleta 2, que apresenta o maior coeficiente. Ainda, o efeito aleatório associado a indivíduo, b_k, tem desvio padrão de 0,6204 e para o efeito aleatório de altura dentro de ponto, $b_{(j|k)}$, o desvio padrão é de 0,2364, o que representa cerca de 45% de toda a variabilidade incluindo o desvio residual.

Em relação ao preditor de σ, tem-se que as estimativas dos coeficientes de regressão associadas a talhão não foram significativas, apresentando valores muito baixos, ou seja,

Figura 3.20 – *Worm plot* do modelo final - Binomial Negativa Tipo II.



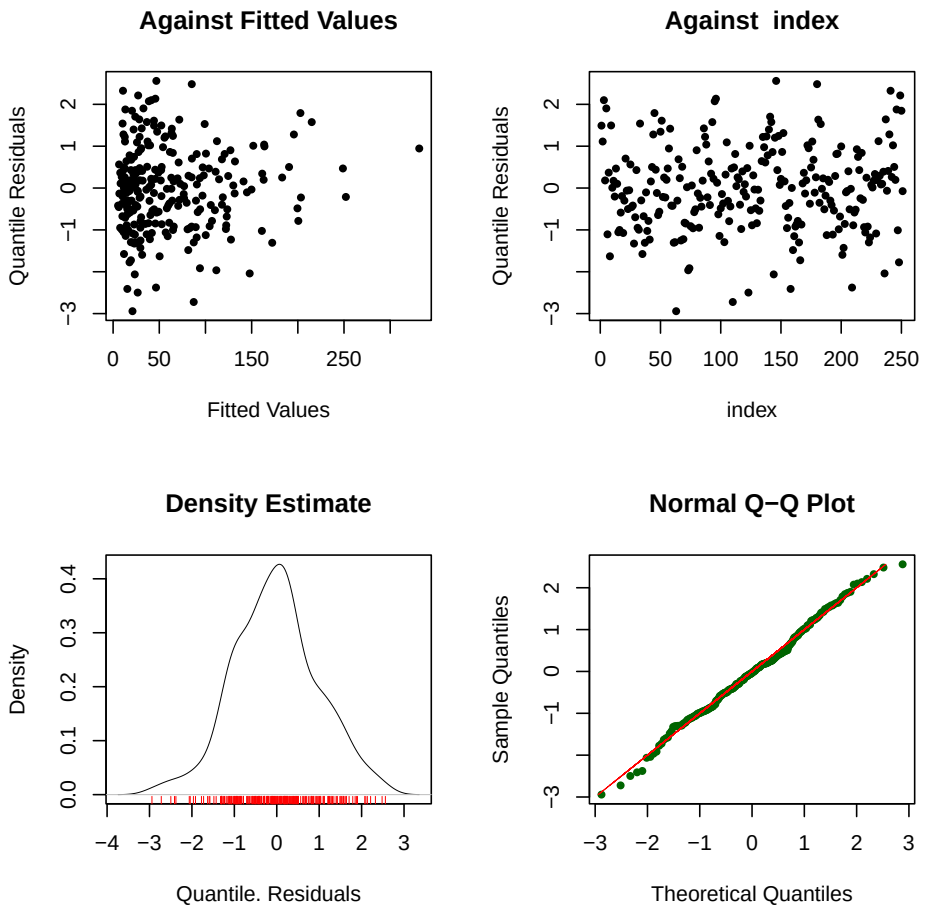
pelo modelo, a variabilidade é similar em cada talhão. Para o fator coleta, todos os coeficientes estimados foram positivos e o maior deles atribui-se a coleta 2, como também observado na análise descritiva, neste momento, ocorreram picos de contagens acima do padrão geral dos dados, contribuindo para uma dispersão maior.

3.4 Modelagem 3 – Localização 2

Para a modelagem da localização 2, a metodologia para a seleção do modelo é analoga à modelagem anterior para a localização 1. Foram considerados, inicialmente, modelos lineares generalizados mistos com distribuições de probabilidade Binomial Negativa Tipo I e Tipo II, com duas estruturas aleatórias diferentes: ponto e ponto e altura, conforme indica a Tabela 3.24. Neste caso, por meio dos *worm plots* dos modelos (Figura 3.22), observa-se que não há a presença de valores discrepantes e o ajuste considerando apenas a modelagem de μ via preditor linear, como é o caso da classe de modelos MLGM, não foi suficiente para obter um bom ajuste. Com isso, faz-se necessário utilizar da flexibilidade da classe de modelos GAMLSS para modelar o preditor linear para σ e testar os modelos indicados na Tabela 3.25.

Na Figura 3.23, por sua vez, há indícios de que o ajuste foi suficientemente adequado

Figura 3.21 – Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II - modelo final.



apenas com a inclusão do preditor linear de σ. Todos os modelos testados se comportaram bem e os pressupostos foram satisfeitos, mesmo com estruturas aleatórias diferentes. Visualmente, com a distribuição de probabilidade Binomial Negativa Tipo I, a estrutura apenas com ponto como fator de efeito aleatório teve um resultado melhor (Modelo 11.1), e com a Binomial Negativa Tipo II, a estrutura com ponto e altura como aleatório foi superior. Apesar disso, tanto nos modelos iniciais quanto nos modelos completos, os critérios AIC associados a ponto e altura como aleatórios foram altamente superiores, corroborado pelo teste da razão de verossimilhança apresentado na Tabela 3.26. Além disso, no sentido prático, incluir altura como fator de efeito aleatório é desejado pois induz uma correlação natural entre as medidas realizadas no mesmo ponto, em alturas diferentes, o que é esperado.

Confrontando os melhores modelos de cada distribuição (Figura 3.24), percebe-se que os resíduos se comportam de maneira muito similar. Desta forma, considerando que o AIC do modelo 2 é melhor e a estrutura com ponto e altura é preferida, como comentado anteri-

Tabela 3.23 – Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II - modelo final.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($> \|t\|$)
			μ		
(Intercepto)	$\boldsymbol{\beta_0}$	**3,36**	**0,08**	**41,74**	**< 0,001**
Talhão 2	α_2	-0,07	0,07	-0,95	0,34
Talhão 3	$\boldsymbol{\alpha_3}$	**-0,99**	**0,10**	**-9,90**	**< 0,001**
Coleta 2	$\boldsymbol{\gamma_2}$	**1,36**	**0,11**	**12,93**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,92**	**0,10**	**9,13**	**< 0,001**
Coleta 4	$\boldsymbol{\gamma_4}$	**0,85**	**0,11**	**7,43**	**< 0,001**
Coleta 5	γ_5	0,23	0,13	1,71	0,09
Coleta 6	$\boldsymbol{\gamma_6}$	**0,65**	**0,11**	**6,10**	**< 0,001**
Coleta 7	$\boldsymbol{\gamma_7}$	**0,55**	**0,12**	**4,52**	**< 0,001**
			σ		
(Intercepto)	$\boldsymbol{\beta_0}$	**1,45**	**0,32**	**4,59**	**< 0,001**
Talhão 2	α_2	-0,03	0,32	-0,08	0,94
Talhão 3	α_3	0,17	0,29	0,59	0,55
Coleta 2	$\boldsymbol{\gamma_2}$	**1,51**	**0,41**	**3,71**	**< 0,001**
Coleta 3	γ_3	0,51	0,46	1,11	0,27
Coleta 4	$\boldsymbol{\gamma_4}$	**1,30**	**0,43**	**3,02**	**< 0,005**
Coleta 5	$\boldsymbol{\gamma_5}$	**1,05**	**0,40**	**2,59**	**0,01**
Coleta 6	$\boldsymbol{\gamma_6}$	**0,97**	**0,40**	**2,45**	**0,02**
Coleta 7	$\boldsymbol{\gamma_7}$	**1,07**	**0,43**	**2,46**	**0,01**

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,3849	0,6204
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,0559	0,2364
Resíduo	ϵ	1,1029	1,0502

Tabela 3.24 – Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
11	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	1970,44
11.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2033,96
12	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	-	1985,39
12.1		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	-	2054,26

ormente, o modelo 2 será selecionado como candidato a modelo final. Pelo *plot* do modelo (Figura 3.25), é possível ver que os resíduos simulados estão aleatoriamente distribuidos,

Figura 3.22 – *Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ.

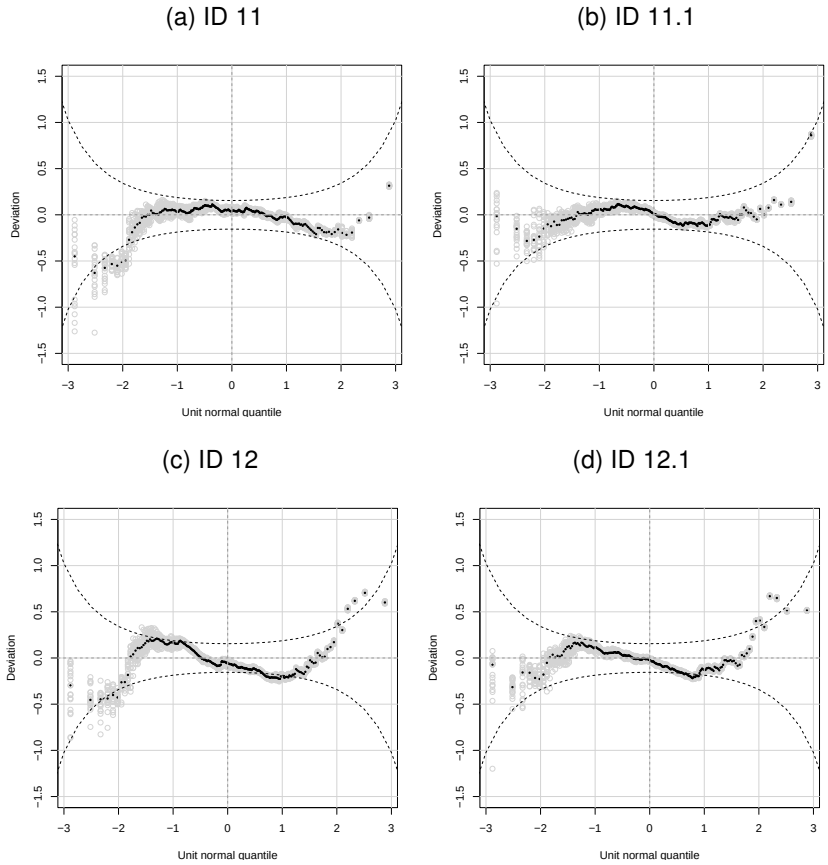
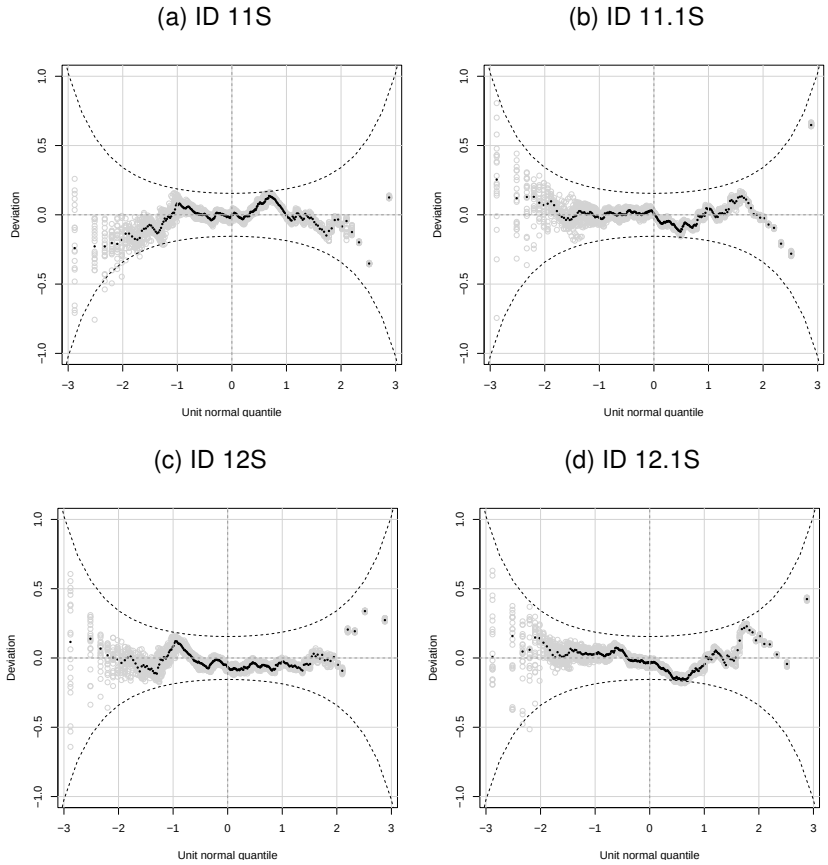


Tabela 3.25 – Modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
11S	Binomial Negativa Tipo I	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	1936,81
11.1S		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	2002,41
12S	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	1960,69
12.1S		$(\beta_0 + b_k) + \alpha_i + \theta_j + \gamma_t$	$\alpha_i + \theta_j + \gamma_t$	2013,40

sendo que os resíduos *versus* valores ajustados estão altamente concentrados entre 0 e 80, no eixo x, com alguns pontos além dessa faixa, como esperado dado o comportamento geral da distribuição dos dados. Quanto a normalidade, cujas medidas descritivas dos resíduos simulados se encontram na Tabela 3.27, o pressuposto é satisfeito e a distribuição dos dados se assemelha a uma normal padrão. No entanto, é possível observar que há um certo desvio em relação a assimetria e curtose o qual se espera corrigir, para o modelo final, após o refinamento.

Figura 3.23 – *Worm plot* dos modelos ajustados com distribuições Binomial Negativa Tipo I e II, com preditor linear para μ e σ.



Em comparação com as modelagens conjunta e da localização 1 separadamente, tem-se que a distribuição de probabilidade Binomial Negativa Tipo II, na forma de $Var(Y) = \mu + \sigma\mu$, se mostrou mais adequada para modelar os dados separadamente, tanto da localização 1 quanto da localização 2, considerando a metodologia aplicada neste trabalho. Por outro lado, a distribuição Binomial Negativa Tipo I, cuja relação média-variância se dá na forma quadrática, foi mais adequada para modelar os dados conjuntamente. Vale destacar que existe uma infinidade de possibilidades e modelos diferentes a serem considerados e testados, e a escolha por um "melhor" modelo, pela semântica, não é possível. Nesse sentido, define-se uma estratégia com base em critérios pré-estabelecidos para encontrar um modelo parcimonioso e que satisfaça os objetivos do pesquisador, cuja capacidade preditiva seja suficientemente boa.

Os coeficientes de regressão apresentados na Tabela 3.28 apontam para uma alta significância dos fatores na predição da média, sendo que apenas as coletas 4 e 7 não foram significativas em relação ao seu *baseline*. Já para a predição do desvio, apenas talhão 3, altura 7 e coleta 7 foram significativos. A partir do modelo completo, foram testados os

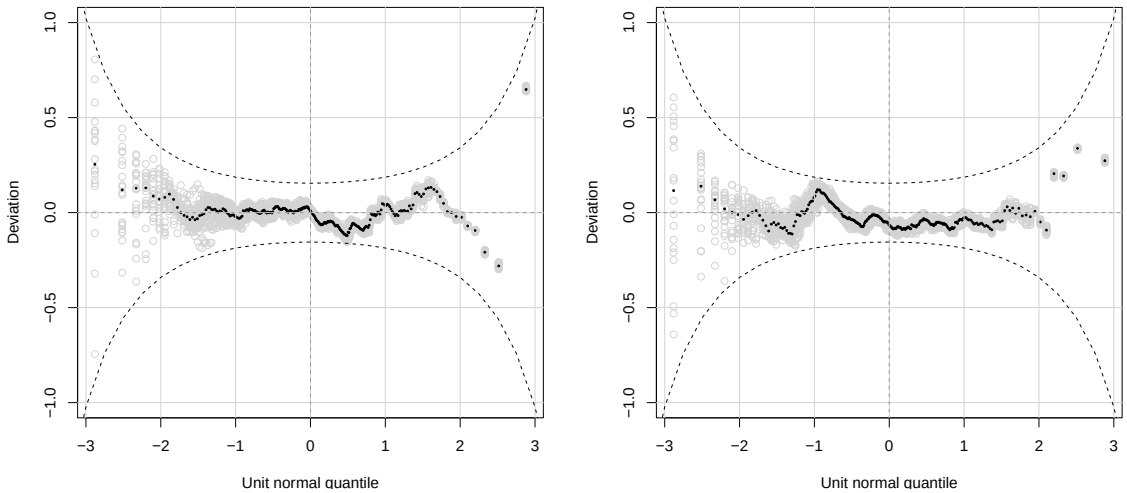
Tabela 3.26 – Teste da razão de verossimilhança entre os modelos com diferentes estruturas aleatórias.

Modelo 11.1S (Nulo):	Deviance = 1939,69	com 31,3623 graus de liberdade
Modelo 11S (Alternativo):	Deviance = 1833,80	com 51,5041 graus de liberdade
LRT = 105,8839 com 20,1418 graus de liberdade		
p-valor < 0,0001		

Modelo 12.1S (Nulo):	Deviance = 1952,22	com 30,5906 graus de liberdade
Modelo 12S (Alternativo):	Deviance = 1866,44	com 47,1283 graus de liberdade
LRT = 85,7830 com 16,5377 graus de liberdade		
p-valor < 0,0001		

Figura 3.24 – Modelos com distribuição de probabilidade Binomial Negativa Tipo I e Tipo II com melhor desempenho.

(a) ID 11.1S (b) ID 12S



modelos exibidos na Tabela 3.29 com a remoção de alguns fatores. Pode-se perceber que, com exceção dos dois últimos modelos na tabela, a remoção dos fatores ocasiona uma diminuição no AIC, embora a diferença entre eles seja bem pequena.

Não por coincidência, uma vez que foi observado na análise descritiva que o comportamento da distribuição se difere dentre os níveis dos fatores, a remoção do fator altura e de talhão e altura simultaneamente no preditor de σ foi responsável por perda na qualidade do ajuste, principalmente com relação à dispersão. Curiosamente, a remoção apenas do fator talhão, ainda que o modelo tenha um critério menor, ocasionou um ajuste inferior ao modelo com a remoção apenas do fator altura.

Em relação à média, foi observado que a remoção apenas do fator altura do preditor

Figura 3.25 – Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II.

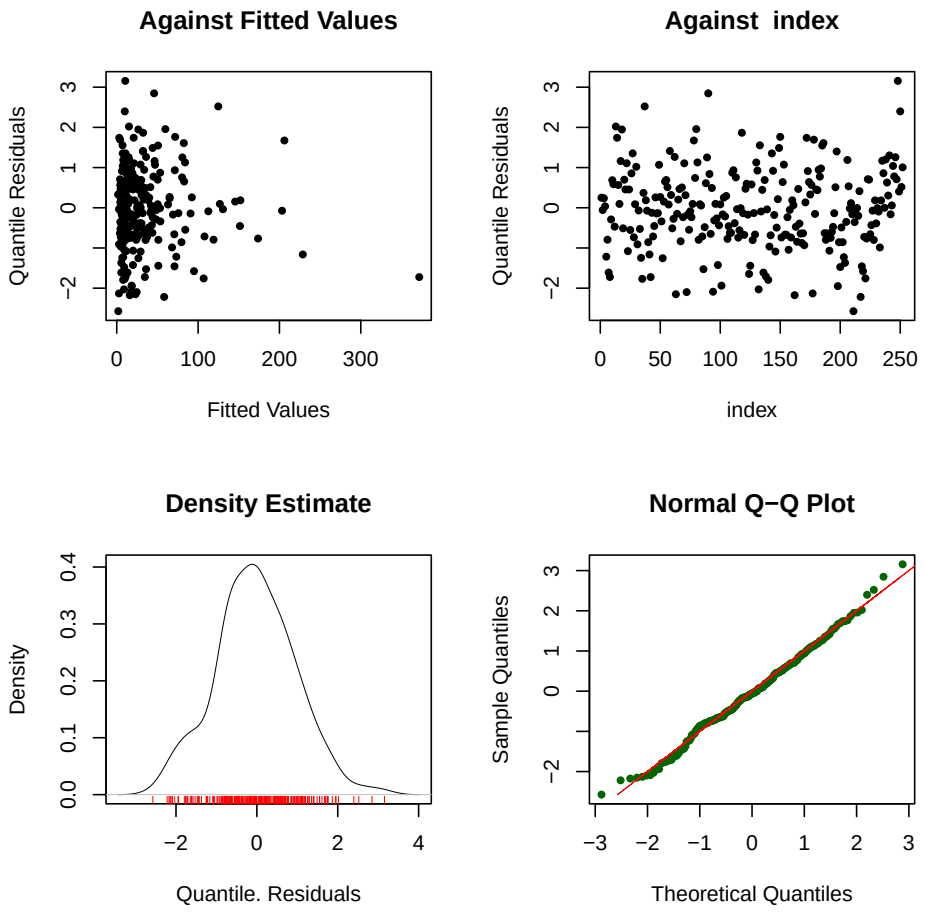


Tabela 3.27 – Medidas descritivas dos resíduos simulados do modelo Binomial Negativa Tipo II.

Média	-0,031
Variância	0,996
Coeficiente de Assimetria	0,119
Coeficiente de Curtose	3,12

Tabela 3.28 – Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($>\|t\|$)
			μ		
(Intercepto)	$\boldsymbol{\beta_0}$	**2,91**	**0,13**	**23,18**	**< 0,001**
Talhão 2	$\boldsymbol{\alpha_2}$	**-0,95**	**0,09**	**-11,13**	**< 0,001**
Talhão 3	$\boldsymbol{\alpha_3}$	**-1,72**	**0,09**	**-18,18**	**< 0,001**
Altura 4,5	$\boldsymbol{\theta_2}$	**0,61**	**0,08**	**7,76**	**< 0,001**
Altura 7	$\boldsymbol{\theta_3}$	**0,53**	**0,11**	**5,00**	**< 0,001**
Coleta 2	$\boldsymbol{\gamma_2}$	**1,14**	**0,13**	**8,67**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,55**	**0,13**	**4,19**	**< 0,001**
Coleta 4	γ_4	0,09	0,13	0,72	0,47
Coleta 5	$\boldsymbol{\gamma_5}$	**-0,51**	**0,15**	**-3,39**	**< 0,001**
Coleta 6	$\boldsymbol{\gamma_6}$	**-0,37**	**0,15**	**-2,48**	**< 0,05**
Coleta 7	γ_7	0,38	0,21	1,76	0,08
			σ		
(Intercepto)	$\boldsymbol{\beta_0}$	**1,81**	**0,41**	**4,43**	**< 0,001**
Talhão 2	α_2	-0,05	0,30	-0,16	0,87
Talhão 3	$\boldsymbol{\alpha_3}$	**-0,79**	**0,34**	**-2,31**	**< 0,05**
Altura 4,5	θ_2	0,00	0,30	-0,01	0,99
Altura 7	$\boldsymbol{\theta_3}$	**1,29**	**0,30**	**4,32**	**< 0,001**
Coleta 2	γ_2	0,58	0,43	1,34	0,18
Coleta 3	γ_3	-0,19	0,43	-0,45	0,66
Coleta 4	γ_4	-0,92	0,49	-1,85	0,07
Coleta 5	γ_5	-0,80	0,48	-1,65	0,10
Coleta 6	γ_6	-0,65	0,46	-1,40	0,16
Coleta 7	$\boldsymbol{\gamma_7}$	**1,34**	**0,45**	**3,00**	**< 0,005**

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,0587	0,2423
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,1596	0,3995
Resíduo	ϵ	1,1157	1,0563

foi suficiente para obter um bom ajuste, lembrando que as conclusões consideradas aqui foram tomadas em conjunto com a inspeção gráfica dos *worm plots* dos modelos. Desta forma, o modelo que melhor se ajustou aos dados contempla, para o preditor de μ, talhão e coleta como fatores de efeitos fixos e ponto e altura como fatores de efeitos aleatórios e, para o preditor de σ, altura e coleta como fatores de efeitos fixos.

Tabela 3.29 – Modelos testados com a distribuição Binomial Negativa Tipo II.

ID	Distribuição	$g_1(\mu)$	$g_2(\sigma)$	AIC
12S	Binomial Negativa Tipo II	$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	1960,69
		$(\beta_0 + b_{(j\|k)} + b_k) + \theta_j + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	1958,06
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	1958,18
		$(\beta_0 + b_{(j\|k)} + b_k) + \gamma_t$	$\beta_0 + \alpha_i + \theta_j + \gamma_t$	1955,70
		$\boldsymbol{(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t}$	$\boldsymbol{\beta_0 + \theta_j + \gamma_t}$	**1959,54**
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \alpha_i + \gamma_t$	1973,52
		$(\beta_0 + b_{(j\|k)} + b_k) + \alpha_i + \gamma_t$	$\beta_0 + \gamma_t$	1974,05

Figura 3.26 – *Worm plot* do modelo final - Binomial Negativa Tipo II.

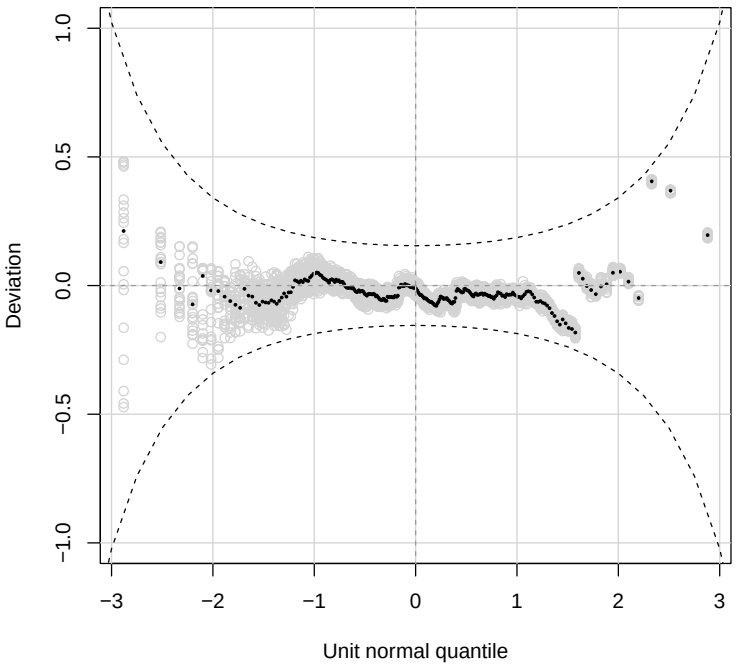
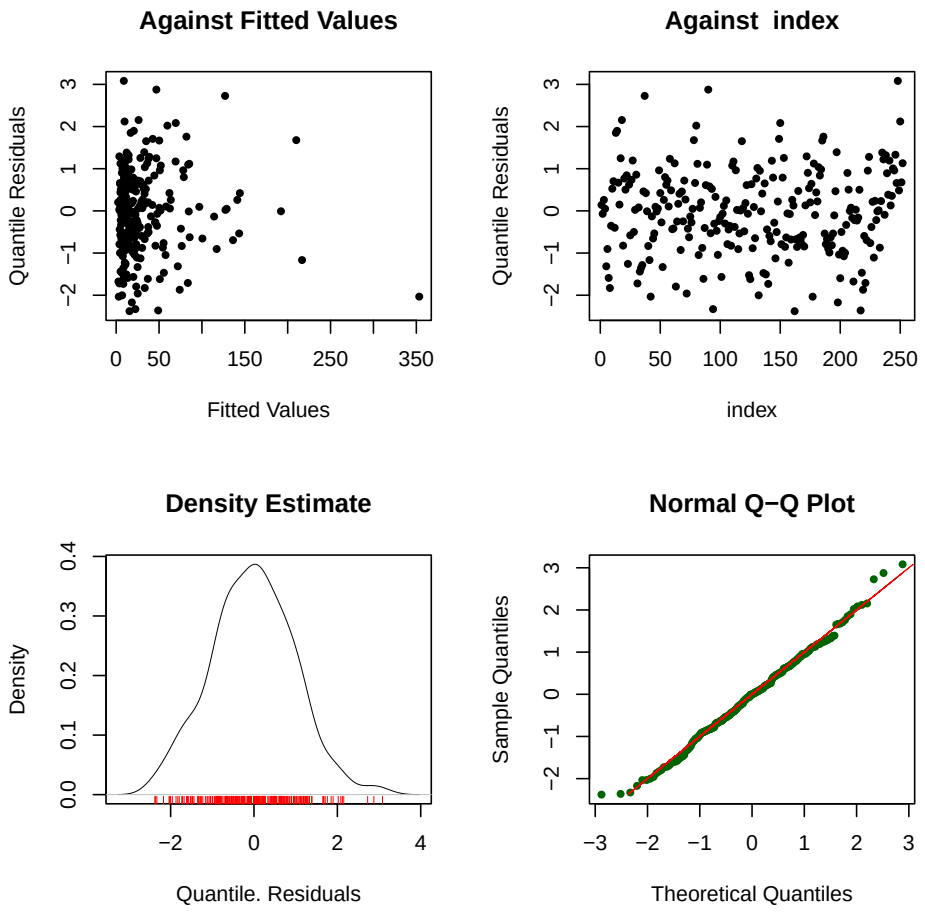


Tabela 3.30 – Medidas descritivas dos resíduos simulados do modelo final - Binomial Negativa Tipo II.

Média	-0,03
Variância	1,004
Coeficiente de Assimetria	0,115
Coeficiente de Curtose	3,03

A partir do modelo selecionado como modelo final, verifica-se pelas medidas descritivas dos resíduos simulados, na Tabela 3.30, que o modelo foi capaz de melhorar o coeficiente de curtose trazendo a medida para próximo de 3, que é a referência de uma distribuição

Figura 3.27 – Gráfico de diagnóstico dos resíduos simulados da distribuição Binomial Negativa Tipo II - modelo final.



mesocúrtica tal qual a distribuição normal. Por outro lado, o coeficiente de assimetria se manteve praticamente estável e acima do valor de referência.

Apesar dessas considerações, pode-se considerar o modelo encontrado como um bom candidato para o ajuste aos dados, visto que os pressupostos foram todos satisfeitos e o desvio observado no *worm plot* é leve, ocorrendo com apenas alguns pontos na cauda, assim como no *QQ Plot* (Figura 3.27), além de ter todos os pontos dentro da banda de confiança. Uma outra alternativa que pode ser considerada é buscar por distribuições de probabilidade que possibilitem o ajuste de dados assimétricos, especialmente dentro da classe de modelos GAMLSS que não necessita pertencer à família exponencial de distribuições. Considerando o modelo final encontrado, portanto, admite-se as funções de

Tabela 3.31 – Estimativas dos coeficientes de regressão e parâmetros de variância da distribuição Binomial Negativa Tipo II - modelo final.

Efeitos Fixos		Estimativa	Erro Padrão	Valor-t	P($>\|t\|$)
			μ		
(Intercepto)	$\boldsymbol{\beta_0}$	**3,07**	**0,10**	**29,97**	**< 0,001**
Talhão 2	$\boldsymbol{\alpha_2}$	**-0,95**	**0,08**	**-12,57**	**< 0,001**
Talhão 3	$\boldsymbol{\alpha_3}$	**-1,58**	**0,09**	**-16,93**	**< 0,001**
Coleta 2	$\boldsymbol{\gamma_2}$	**1,10**	**0,12**	**9,15**	**< 0,001**
Coleta 3	$\boldsymbol{\gamma_3}$	**0,58**	**0,12**	**4,69**	**< 0,001**
Coleta 4	γ_4	0,08	0,12	0,71	0,48
Coleta 5	$\boldsymbol{\gamma_5}$	**-0,52**	**0,14**	**-3,83**	**< 0,001**
Coleta 6	$\boldsymbol{\gamma_6}$	**-0,36**	**0,14**	**-2,60**	**< 0,05**
Coleta 7	γ_7	0,15	0,17	0,86	0,39
			σ		
(Intercepto)	$\boldsymbol{\beta_0}$	**1,78**	**0,34**	**5,31**	**< 0,001**
Altura 4,5	θ_2	-0,23	0,27	-0,84	0,40
Altura 7	$\boldsymbol{\theta_3}$	**1,02**	**0,26**	**3,87**	**< 0,001**
Coleta 2	γ_2	0,37	0,40	0,92	0,36
Coleta 3	γ_3	0,01	0,41	0,03	0,98
Coleta 4	$\boldsymbol{\gamma_4}$	**-1,03**	**0,48**	**-2,17**	**< 0,05**
Coleta 5	γ_5	-0,80	0,45	-1,76	0,08
Coleta 6	γ_6	-0,53	0,44	-1,20	0,23
Coleta 7	$\boldsymbol{\gamma_7}$	**0,97**	**0,40**	**2,44**	**< 0,05**

Efeitos Aleatórios		Variância	D. Padrão
Ponto (Intercepto)	b_k	0,0241	0,1553
Altura em Ponto (Intercepto)	$b_{(j\|k)}$	0,2689	0,5186
Resíduo	ϵ	1,1246	1,0605

ligação:

$$
\begin{aligned}
log(\hat{\mu}_{ijkt}) = \hat{\eta}^{\mu}_{ijkt} &= (3,07 + \hat{b}_k + \hat{b}_{(j|k)}) - 0,95\, Talhão\ 2 - 1,58\, Talhão\ 3 + \\
&+\ 1,10\, Coleta\ 2 + 0,58\, Coleta\ 3 + 0,08\, Coleta\ 4 + \\
&-\ 0,52\, Coleta\ 5 - 0,36\, Coleta\ 6 + 0,15\, Coleta\ 7 \\
log(\hat{\sigma}_{ijkt}) = \hat{\eta}^{\sigma}_{ijkt} &= 1,78 - 0,23\, Altura\ 4,5 + 1,02\, Altura\ 7 + 0,37\, Coleta\ 2 + \\
&+\ 0,01\, Coleta\ 3 - 1,03\, Coleta\ 4 - 0,80\, Coleta\ 5 + \\
&-\ 0,53\, Coleta\ 6 + 0,97\, Coleta\ 7.
\end{aligned}
\tag{3.4.1}
$$

A partir da Equação (3.4.1) referente ao modelo final, para o preditor de μ, apenas as coletas 4 e 7 não foram significativas em relação ao *baseline*. Os talhões 2 e 3 foram sig-

nificativos e possuem coeficientes de regressão negativos, ou seja, a contagem média de insetos nesses talhões é menor, sendo no talhão 3 a menor contagem média. Com exceção das coletas 5 e 6, todos os demais coeficientes associados a esse fator foram positivos, sendo a contagem média maior na coleta 2. Entre as coletas 5 e 6, cujas estimativas são negativas, a coleta 5 possui contagem média menor de insetos mas parece não diferir da coleta 6, visto que o valor desses coeficientes são próximos. Embora o fator altura tenha apresentado alta significância no modelo inicial (Tabela 3.28), a sua remoção do modelo não acarreta perda na qualidade do ajuste. Além disso, o efeito aleatório de altura dentro de ponto apresenta um desvio padrão maior no modelo final, de 0,5186 contra 0,3995 do modelo inicial, incorporando essa variabilidade devido a altura no modelo. O desvio padrão do efeito aleatório a nível de indivíduo foi de 0,1553, e neste caso, a variabilidade devido a altura dentro de ponto é maior do que a variabilidade entre indivíduos, diferentemente do que ocorre na modelagem conjunta e da localização 1. Juntos, os desvios associados aos efeitos aleatórios representam 38,85% do total.

Em relação ao preditor de σ, o menor desvio na contagem ocorre na altura 4,5 e o maior, na altura 7. Para o fator coleta, alguns coeficientes são negativos e outros positivos, indicando que a variabilidade na coleta 1 e altura 2,5 não é a maior e nem a menor. O maior desvio padrão, pelo modelo, ocorre na altura 7 e coleta 7 enquanto o menor ocorre na altura 4,5 e coleta 4.

Capítulo 4

Conclusão

A partir da modelagem conjunta, foi possível observar diferenças na contagem média de insetos entre os talhões que pertencem a diferentes localizações. Além disso, o efeito dos fatores na resposta pode se manifestar de formas diferentes em cada local, assim como o comportamento da distribuição da contagem de insetos, conforme explorado na análise descritiva dos dados. Desta forma, procedeu-se com a modelagem de forma separada em cada localização com vistas a incorporar tais particularidades de cada uma delas e encontrar um modelo mais adequado.

É possível observar diferenças e similaridades entre os modelos ajustados, sendo que a principal diferença está no fato de que a distribuição de probabilidade Binomial Negativa Tipo I, na forma quadrática da relação média-variância, foi mais adequada para modelar o banco de dados completo, enquanto para os modelos separados, a Binomial Negativa Tipo II obteve melhor desempenho. Quanto as similaridades entre os modelos, para ambas localizações, a distribuição Binomial Negativa Tipo II proporcionou um ajuste adequado, embora a distribuição da contagem de insetos na localização 2 tenha apresentado um comportamento mais concentrado e menos disperso em relação a localização 1. Nos três modelos, para a média, a estrutura do preditor linear foi a mesma, com talhão e coleta como fatores de efeitos fixos e ponto e altura como fatores de efeitos aleatórios. Para o preditor do desvio, na modelagem conjunta e da localização 1, os fatores de efeitos fixos foram talhão e coleta, enquanto para a localização 2, apenas os fatores altura e coleta foram importantes para o modelo.

Com relação a metodologia utilizada, percebe-se que mesmo a classe de modelos GAMLSS permitindo a modelagem de distribuições que não necessitam pertencer à família exponencial, generalizando tal suposição, curiosamente, duas das três melhores dentre

as distribuições testadas foram Binomiais Negativas que podem ser escritas sob a forma da família exponencial. Deste modo, inicialmente, foi possível ajustar um modelo linear generalizado misto incluindo o preditor linear apenas para a média e avaliar o ajuste, no entanto, o modelo não foi suficiente para obter um bom ajuste. A modelagem de σ via preditor linear utilizando a flexibilidade da classe de modelos GAMLSS se mostrou fundamental para melhorar consideravelmente o ajuste aos dados, o que era esperado uma vez que foi possível constatar, na análise exploratória dos dados, que a variabilidade da contagem de insetos não é constante sob os níveis dos fatores.

Referências

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

ALMEIDA, E. P. d.; JANEIRO, V.; GUEDES, T. A.; MULATI, F.; CARNEIRO, J. W. P.; NUNES, W. M. d. C. Modeling citrus huanglongbing data using a zero-inflated negative binomial distribution. *Acta Scientiarum. Agronomy*, SciELO Brasil, v. 38, n. 3, p. 299–306, 2016.

ALVES, G. R. *Efeitos de variedades cítricas sobre o vetor das bactérias associadas ao Huanglongbing (HLB) Diaphorina citri Kuwayama, 1908 (Hemiptera: Liviidae)*. Tese (Doutorado) — Universidade de São Paulo, 2017.

BATISTA, D. T. *Modelos para dados de contagem não equidispersos com aplicação à ecologia e em estudos longitudinais*. Tese (Doutorado) — Universidade de São Paulo, 2020.

BOVÉ, J. M. Huanglongbing: a destructive, newly-emerging, century-old disease of citrus. *Journal of Plant Pathology*, v. 88, p. 7–37, 2006.

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

CARVALHO, F. J. Modelos lineares generalizados na agronomia: análise de dados binomiais e de contagem, zeros inflacionados e enfoque bayesiano. Universidade Federal de Uberlândia, 2019.

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

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. *Modelos lineares generalizados e extensões*. [S.l.]: Piracicaba: USP, 2008.

COSTA, S. C. da. *Modelos lineares generalizados mistos para dados longitudinais*. Tese (Doutorado) — Universidade de São Paulo, 2003.

DIGGLE, P. J.; HEAGERTY, P.; LIANG, K.-Y.; ZEGER, S. L. *Analysis of longitudinal data*. [S.l.]: Oxford University Press, 2002.

DOBSON, A. J. *An Introduction to Generalized Linear Models*. [S.l.]: Chapman & Hall/CRC, 2002.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, p. 236–244, 1996.

FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. *Applied longitudinal analysis*. [S.l.]: John Wiley & Sons, 2012. v. 998.

FOX, J. *Applied regression analysis and generalized linear models*. [S.l.]: Sage Publications, 2015.

GALDINO, M. V. *Modelos lineares generalizados mistos e equações de estimação generalizadas para dados binário aplicados em anestesiologia veterinária*. Dissertação (Mestrado) — Universidade Estadual Paulista (UNESP), 2015.

GREGO, S. *Ajuste de modelos lineares generalizados geoestatísticos em dados de contagem com diferentes abordagens agroclimáticas*. Tese (Doutorado) — Universidade de São Paulo, 2019.

HALBERT, S. E.; MANJUNATH, K. L. Asian citrus psyllids (sternorrhyncha: Psyllidae) and greening disease of citrus: a literature review and assessment of risk in florida. *Florida Entomologist*, v. 87, n. 3, p. 330–353, 2004.

HARDIN, J. W.; HILBE, J. M. *Generalized linear models and extensions*. [S.l.]: Stata press, 2018.

HARVILLE, D. A. Bayesian inference for variance components using only error contrasts. *Biometrika*, Oxford University Press, v. 61, n. 2, p. 383–385, 1974.

HESS, A. F.; CIANORSCHI, L.; SILVESTRE, R.; SCARIOT, R.; RICKEN, P. Aplicação dos modelos lineares generalizados para estimativa do crescimento em altura. *Pesquisa Florestal Brasileira*, v. 35, n. 84, p. 427–433, 2015.

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

MAYR, A.; FENSKE, N.; HOFNER, B.; KNEIB, T.; SCHMID, M. Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 61, n. 3, p. 403–427, 2012.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: Chapman & Hall/CRC, 1989. v. 37.

MCCULLOCH, C. E.; SEARLE, S. R. *Generalized, Linear, and Mixed Models*. [S.l.]: Wiley-Interscience, 2001.

NAVA, D. T. Modelos lineares generalizados e processos pontuais em análise espacial de dados agrícolas. Universidade Estadual do Oeste do Paraná, 2018.

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

OLIVEIRA, E. S. C.; ANDRADE, H. A. Padronização da taxa de captura para o espadarte (xiphias gladius) usando efeito fixo e aleatório. *Brazilian Journal of Agricultural Sciences/Revista Brasileira de Ciências Agrárias*, v. 14, n. 1, 2019.

OLIVEIRA, R. A. d.; BATISTELA, G. C.; SIMÕES, D.; PADOVANI, C. R. Aplicação de modelos lineares de efeitos mistos para avaliar a densidade básica da madeira de duas espécies e um híbrido de eucalyptus. v. 49, 03 2021.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, Oxford University Press, v. 58, n. 3, p. 545–554, 1971.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2013.

PINHEIRO, J.; BATES, D. *Mixed-effects models in S and S-PLUS*. [S.l.]: Springer Science & Business Media, 2006.

PITT, D.; TRÜCK, S.; H., R. van den; WONG, W. W. Modeling risks from natural hazards with generalized additive models for location, scale and shape. *Journal of Environmental Management*, Elsevier, v. 275, p. 111075, 2020.

RIGBY, R. A.; STASINOPOULOS, D. M. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, Springer, v. 6, n. 1, p. 57–65, 1996.

RIGBY, R. A.; STASINOPOULOS, M. D. 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.

SÁ, A. C. L.; TURKMAN, M. A. A.; PEREIRA, J. Exploring fire incidence in portugal using generalized additive models for location, scale and shape (gamlss). *Modeling Earth Systems and Environment*, Springer, v. 4, n. 1, p. 199–220, 2018.

SANTOS, D. A. S.; AZEVEDO, P. V. d.; OLINDA, R. A. d.; SANTOS, C. A. C. d.; SOUZA, A. d.; SETTE, D. M.; SOUZA, P. M. d. A relação das variáveis climáticas na prevalência de infecção respiratória aguda em crianças menores de dois anos em rondonópolis-mt, brasil. *Ciência & saúde coletiva*, SciELO Public Health, v. 22, p. 3711–3722, 2017.

SERINALDI, F. Distributional modeling and short-term forecasting of electricity prices by generalized additive models for location, scale and shape. *Energy Economics*, Elsevier, v. 33, n. 6, p. 1216–1226, 2011.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, p. 1–46, 2008.

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. D. *Flexible regression and smoothing: using GAMLSS in R*. [S.l.]: CRC Press, 2017.

TENNANT, P. F.; ROBINSON, D.; FISHER, L.; BENNETT, S.-M.; HUTTON, D.; COATES-BECKFORD, P.; LAUGHLIN, W. M. Diseases and pests of citrus (*Citrus* spp.). *Tree and Forestry Science and Biotechnology*, v. 3, p. 81–107, 2009.

TUAN, T. S.; VENÂNCIO, T. S.; NASCIMENTO, L. F. C. Efeitos da exposição a poluentes do ar no infarto agudo do miocárdio, segundo gêneros. *Arquivos Brasileiros de Cardiologia*, SciELO Brasil, v. 107, n. 3, p. 216–222, 2016.

VALIM, W. B. S. *Monitoramento de migração de Diaphorina citri em pomar comercial de laranja*. Dissertação (Mestrado) — Universidade Estadual de Maringá, 2021. Em fase de elaboração.

VERBEKE, G.; MOLENBERGHS, G. *Linear mixed models for longitudinal data*. [S.l.]: Springer Science & Business Media, 2000.

WEST, B. T.; WELCH, K. B.; GALECKI, A. T. *Linear mixed models: A practical guide using statistical software*. [S.l.]: Taylor & Francis, 2007.

WOLFINGER, R. Covariance structure selection in general mixed models. *Communications in statistics-Simulation and computation*, Taylor & Francis, v. 22, n. 4, p. 1079–1106, 1993.