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

Sthefany Caroline Volpato

Comparações de Distribuições de Probabilidade na Análise à Resistência ao Cancro Cítrico.

Maringá – Paraná
2021

Sthefany Caroline Volpato

Comparações de Distribuições de Probabilidade na Análise à Resistência ao Cancro Cítrico.

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.

Orientadora: Profª. Drª. Terezinha Aparecida Guedes
Coorientador: Prof. Dr. Vanderly Janeiro

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá – Paraná

2021

Sthefany Caroline Volpato

Comparações de Distribuições de Probabilidade na Análise à Resistência ao Cancro Cítrico.

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

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

Prof. Dr. Rodrigo Rosseto Pescim
Universidade Estadual de Londrina – Depto. Estatística/UEL

Prof. Dr. Willian Luís de Oliveira
Universidade Estadual de Maringá – PBE/UEM

Maringá, 21 de Maio de 2021.

Este trabalho é dedicado à Deus, à minha mãe e aos meus irmãos. Sem vocês eu nada seria e nada conseguiria. Obrigada!

Agradecimentos

Primeiramente, agradeço à Deus, por toda força e coragem diária.

À minha família, por todo amor, apoio, dedicação e confiança. Além de dar todo o suporte necessário para que meus estudos fossem concluídos, incentivando todas as decisões tomadas nesse período e por sempre entenderem os momentos difíceis e de ausência.

À minha orientadora Terezinha Aparecida Guedes, que dividindo todo seu conhecimento permitiu que eu me interessasse ainda mais pela área da estatística. Um exemplo de profissional a ser seguido, com um coração enorme e que terá a minha gratidão eterna. Agradeço ainda, meu coorientador Vanderly Janeiro, por sempre estar disposto a dar sugestões e indicar melhorias neste trabalho.

Aos amigos adquiridos no mestrado e também as amizades que mantenho fora da universidade, em especial, Bruna, André, Breno, Rafaela, Andressa e Matheus, que tornaram os dias de estudos mais leves e alegres.

Aos professores que tive durante a graduação em matemática e também aos professores do departamento de estatística que pude conhecer durante o mestrado e que com certeza agregaram muito à minha formação como aluna, professora e ser humano.

Á CAPES por todo suporte financeiro durante o período de mestrado. E também à Universidade Estadual de Maringá, pela oportunidade de realizar a graduação e a pós-graduação na instituição.

Resumo

O setor agrário é um grande gerador de pesquisas que necessitam de apoio estatístico principalmente para melhoramento genético de plantas em geral e suscetibilidade à doenças e pragas. Os dados que geralmente são coletados nesse tipo de experimento são de caráter longitudinal com medidas repetidas, ao passo que o intuito do pesquisador é o de verificar o comportamento e evolução da doença ao longo de um período pré determinado, por exemplo. Com base nisso, as metodologias estatísticas a serem utilizadas nesse tipo de modelagem devem conseguir captar o maior número de informações possíveis inerentes ao banco de dados, assim como uma possível dependência intra-indivíduo, já que pode haver correlação entre mensurações feitas na mesma unidade experimental. Nesse sentido, o objetivo deste trabalho foi inicialmente avaliar a resistência de dezesseis genótipos de laranja doce à doença cancro cítrico através da metodologia de modelos lineares generalizados mistos, comparando cinco distribuições de probabilidade assimétricas para verificar qual ajuste seria o mais adequado e posteriormente avaliar a resistência de seis genótipos de laranja doce à mesma doença, porém comparando apenas duas distribuições de probabilidade, que para este caso, são simétricas. O intuito final de ambas as aplicações era verificar o comportamento da variável resposta (diâmetro da lesão), com relação a algumas possíveis variáveis explicativas e concluir qual genótipo seria mais e menos suscetível à doença. Com relação as distribuições de probabilidade utilizadas não houveram diferenças significativas dos ajustes dos dois cenários analisados, sendo que no primeiro o genótipo *Pera 436* foi o que apresenta maior resistência à doença durante todo período de avaliação e o genótipo *Westin 340* se mostra sendo o mais suscetível. Para a segunda aplicação os genótipos mais e menos resistentes são o *Valência* e *Prec Ori*, respectivamente.

Palavras-chave: Dados longitudinais, medidas repetidas, modelo linear generalizado misto, distribuições de probabilidade.

Abstract

Most research areas use statistical support to assist in decision making or to draw conclusions in their studies. The agrarian sector is a major generator of research that needs statistical support mainly for plant breeding in general and susceptibility to diseases and pests. The data that are usually collected in this type of experiment are longitudinal in nature with repeated measures, whereas the researcher's intention is to verify the behavior and evolution of the disease, over a predetermined period, for exemple. Based on this, the statistical methodologies used in those modeling must be able to capture the largest possible number of information inherent to the database, as well as a possible intra-individual dependence, since there may be a correlation between measurements in the same experimental unit. In this sense, the resistance of sixteen sweet orange genotypes to citrus canker disease is evaluated in the first application of this work through the methodology of generalized linear mixed models to compare five asymmetric probability distributions and verify which fit is the most appropriate. In a second application, the resistance of six sweet orange genotypes to the same disease is evaluated to compare only two symmetrical probability distributions. The final purpose of both applications is to verify the behavior of the response variable (lesion diameter) in relation to some possible explanatory variables and to conclude which genotype is more and less susceptible to the disease. Regarding the probability distributions used for the fits, a consensus in all of then in both applications, where in the first one the genotype *Pera 436* present the most resistance to the disease during the entire period of evaluation and the *Westin 340* genotype proves to be the most susceptible. For the second application,the most and least resistant genotypes are *Valencia* and *Prec Ori*, respectively.

Keywords: Longitudinal data, repeated measures, generalized linear mixed model, probability distributions.

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas e siglas

AR	Auto-regressiva
ARH	Auto-regressiva Heterogênea
ARMA	Auto-regressiva de Médias Móveis
BS	Birnbaum Saunders
BSSN	Birnbaum Saunders Skew Normal
CS	Simetria Composta
GA	Gama
GAM	Modelo Aditivo Generalizado
GAMLSS	Modelo Aditivo Generalizado para Locação, Escala e Forma
GLM	Modelo Linear Generalizado
GLMM	Modelo Linear Generalizado Misto
IG	Normal Inversa
LMM	Modelo Linear Misto
LOGNO	Log-normal
NO	Normal
SN	Skew Normal
ST3	Skew-t tipo 3
T	T-Student
TOEP	Toeplitz
UN	Não Estruturada

Sumário

Introdução

A maioria dos pesquisadores das mais diversas áreas de pesquisa utilizam suporte estatístico para auxiliar na tomada de decisão ou tirar conclusões em seus estudos. Para isso, recorrem a métodos de coleta, manuseio e apresentação dos dados coletados, assim como métodos de análise e interpretação. Dessa forma, ferramentas estatísticas se tornam de suma importância em inúmeras áreas do saber.

O setor agrário é um grande gerador de pesquisas que necessitam de apoio estatístico para serem desenvolvidas. A constante necessidade de respostas com relação ao melhoramento genético de plantas em geral, suscetibilidade à doenças e pragas torna esse segmento um grande produtor de dados e usuário dessas análises.

A avaliação da resistência de genótipos à doenças, por exemplo, é uma atividade antiga na pesquisa agrícola e que reflete um grande impacto na produção mundial de alimentos. Os dados que geralmente são coletados nessas pesquisas são longitudinais, visto que muitas das vezes o interesse do pesquisador é investigar como a doença se comporta ao longo de uma dimensão específica (tempo e espaço são exemplos nesse caso). Nesse sentido, para que o mesmo genótipo possa ser avaliado em diferentes condições, medidas repetidas são utilizadas para a obtenção de dados.

Várias abordagens surgiram a fim de extrair o máximo de informação possível e obviamente de forma adequada, para dados longitudinais com medidas repetidas. Littell, Henry e Ammerman (1998), destacam que é preciso verificar a existência de correlação entre as mensurações da mesma unidade amostral em tempos distintos, bem como avaliar se a variabilidade necessita ou não ser modelada.

Dentre as abordagens que já foram propostas na literatura, se sobrepõe o estudo de Laird e Ware (1982) que envolve a especificação de um modelo denominado modelo linear misto (LMM). Essa metodologia inclui além dos efeitos fixos, efeitos aleatórios ao modelo linear padrão, permitindo-se assim modelar de forma adequada a correlação e a variabi-

lidade que possam existir na mesma unidade experimental, ou seja, intra-indivíduo. Um dos pressupostos dessa metodologia é o fato de que a variável de interesse seja normalmente distribuída, ao passo que restringe as análises quando esta assume comportamento assimétrico, por exemplo.

Tendo em vista essa restrição, uma extensão dos LMM foi proposta por Breslow e Clayton (1993), em que consiste em usar os já conhecidos modelos lineares generalizados (GLM) (NELDER; WEDDERBURN, 1972) em conjunto com a modelagem mista, criando-se então uma classe de modelos denominada modelos lineares generalizados mistos (GLMM). O grande diferencial desses modelos é que o pressuposto de normalidade para a variável de interesse não é mais necessário. Assim, pode-se modelar qualquer tipo de variável que tenha distribuição de probabilidade que pertença à família exponencial de distribuições.

Com o intuito de estender ainda mais a quantidade de distribuições de probabilidade que poderiam ser utilizadas na modelagem, de permitir que todos os parâmetros do modelo se relacione à covariáveis existentes e de tornar a forma dos preditores mais flexíveis, estudou-se o modelo aditivo generalizado para locação, escala e forma (GAMLSS). Proposto por Rigby e Stasinopoulos (2001, 2005) e Akantziliotou e Stasinopoulos (2002), e implementado no *Software R*, por meio da função `gamlss`, o GAMLSS relaxa a suposição da variável resposta pertencer a família exponencial. Além disso o GAMLSS, e consequentemente o pacote `gamlss` do R, permitem a introdução de efeitos aleatórios, ou seja, modelagem mista.

Assim, o propósito deste trabalho é apresentar e utilizar a metodologia de modelos lineares generalizados mistos no estudo da suscetibilidade de genótipos de laranja doce à doença cancro cítrico. São apresentadas dois cenários, onde a finalidade é modelar o comportamento da variável resposta (diâmetro da lesão) com relação as variáveis explicativas, tratamento, dias após a inoculação da bactéria (DAI) e folhas, fazendo uso de uma metodologia que consiga representar todas as informações contidas nos dados para que no final da análise possa auxiliar na identificação de qual genótipo dentre os estudados é mais/menos suscetível à doença. Todas as análises estatísticas foram realizadas no *Software* estatístico R versão 3.6.3.

Objetivo

Objetivo Geral

Apresentar um modelo estatístico que seja capaz de captar o máximo de informações contidas nos dados observados da lesão do cancro cítrico, causado pela inoculação da

bactéria *Xanthomonas citri subsp. citri* nas folhas das laranjeiras vistos ao longo do tempo, utilizando a metodologia de modelos lineares generalizados mistos.

Objetivos Específicos

- Explorar o comportamento da variável resposta com relação as variáveis explicativas;
- Descrever por meio do modelo linear generalizado misto o comportamento da variável resposta;
- Efetuar a análise de resíduos dos modelos propostos;
- Verificar se o comportamento da variável resposta diâmetro da lesão pode ser explicado pela mesma distribuição de probabilidade em ambos os experimentos com diferentes variedades de citros.

Capítulo 1

Revisão de Literatura

1.1 Dados Longitudinais e Medidas Repetidas

O termo medidas repetidas é utilizado em uma pesquisa quando os dados são coletados no mesmo indivíduo ou na mesma unidade experimental em mais de uma ocasião. Quando essa coleta é realizada em uma sequência ordenada, os dados são denominados longitudinais.

Os estudos longitudinais são projetados para averiguar mudanças ao longo do tempo em uma certa característica que é medida repetidamente para cada participante do estudo (LAIRD; WARE, 1982). Geralmente apresentam estrutura hierárquica, pois as medidas repetidas são aninhadas dentro do indivíduo. Pode-se então fazer a suposição de que as observações entre os indivíduos sejam independentes e que as aninhadas sejam dependentes e com erros correlacionados.

De acordo com Ware (1985), o grande objetivo da pesquisa longitudinal é caracterizar padrões da variável resposta no tempo, verificando-se os efeitos das covariáveis nesses padrões. Hedeker e Gibbons (2006), destacam que são inúmeras as vantagens dos estudos longitudinais em relação aos transversais (caracterizado pelas "medições"serem feitas em um único momento), por exemplo. Uma delas é o fato de que para alcançar um nível semelhante de poder estatístico, menos sujeitos, ou seja, um tamanho de amostra menor, é necessário em um estudo longitudinal.

A análise de dados longitudinais é bem representada na literatura. O tema foi abordado por importantes autores ao longo da história, como por exemplo, Goldstein (1979) que analisou dados longitudinais educacionais e sociais de mais de 9000 crianças através da abordagem de modelos lineares. O artigo discute e analisa as premissas desses mo-

delos.Também foram estudadas inter-relações complexas ao longo do tempo entre várias variáveis.

Ware (1985) propõe uma abordagem mais flexível aos dados longitudinais que permite a especificação da resposta esperada como uma função linear arbitrária de covariáveis fixas e variáveis no tempo. São discutidos três famílias de modelos para a função de covariância e as ilustrações demonstraram a utilidade da abordagem proposta para a análise longitudinal.

Duncan e Kalton (1987) revisaram algumas técnicas para a análise de dados longitudinais usando modelos de mudança de tempo discretos. Willett, Singer e Martin (1998) apresentaram a análise de estudos longitudinais de desenvolvimento de psicopatia, com ênfase a recomendações metodológicas que fornecem maneiras poderosas de responder suas perguntas de pesquisa sobre mudanças sistemáticas ao longo do tempo no comportamento individual. Diggle et al. (2002) descreveram modelos e métodos estatísticos para a análise de dados longitudinais, com uma forte ênfase em aplicações nas ciências biológicas e da saúde.

Em um trabalho mais recente, Yoon e Jain (2015) analisaram as pontuações de correspondência de impressão digital (similaridade) por meio de modelos estatísticos multinível, com intervalo de tempo entre duas impressões digitais em comparação, idade do sujeito e qualidade da imagem da impressão digital. Registros de impressões digitais longitudinais de 15.597 indivíduos são amostrados a partir de um banco de dados de impressão digital operacional, de modo que cada indivíduo tenha pelo menos cinco registros de 10 impressões em um período mínimo de 5 anos.

Uma das vantagens do estudo com medidas repetidas está pautada no fato de que um indivíduo pode servir como seu próprio controle, permitindo que comparações intra-indivíduos sejam feitas. Littell, Henry e Ammerman (1998), sugeriram que uma maior atenção seja dada para dados desse tipo, devido as possíveis correlações e variações que podem aparecer na resposta para uma mesma unidade amostral. Os experimentos com medidas repetidas nem sempre dão frutos à dados balanceados, visto que muitas das vezes o pesquisador não tem controle sobre como estes são coletados, o que se torna um problema para a modelagem tradicional.

Quando o foco são estudos longitudinais com medidas repetidas sendo utilizados em dados da área agrícola, tem-se como referência o trabalho de Gonçalves-Zuliani (2014), que objetivava avaliar a resistência de 9 genótipos de laranja doce a respeito da doença cancro cítrico, por metodologia de folhas destacadas. Um outro estudo, semelhante a este, foi abordado por Nanami (2017), em que foram avaliados 14 genótipos de citros, com relação à resistência ao cancro cítrico em condição de casa de vegetação.

Khan et al. (2016) realizaram um experimento utilizando o delineamento em blocos causalizados, composto por onze tratamentos repetidos três vezes durante o ano de 2015. O objetivo desse experimento era de verificar o desempenho de diferentes cultivares de laranja doce para a seleção de espécies resistentes ao cancro cítrico.

A suscetibilidade de uma variedade foi avaliada por Graham, Gottwald et al. (1990), onde evidenciou-se que o diâmetro da lesão causada pela doença está também ligada a agressividade da mesma.

Assim, com a importância do tema e sabendo-se que o Brasil é o maior produtor mundial de citros (20% da produção mundial) e o maior exportador mundial de suco de laranja concentrado congelado (50% das exportações mundiais)(TIMMER; GARNSEY; BROADBENT, 2003), tem-se como objetivo estudar a relação da variável resposta diâmetro da lesão com as variáveis explicativas, tratamento, dias após a inoculação da bactéria e folhas, por meio da modelagem mista com a finalidade de encontrar os genótipos mais/menos suscetíveis à doença.

Capítulo 2

Materiais e Métodos

2.1 Materiais

O cancro cítrico é uma das doenças que mais afetam a citricultura mundial. Como é de fácil disseminação, está presente em praticamente todas as regiões de cultivo e leva como responsabilidade grandes prejuízos para os produtores (GOTTWALD; GRAHAM; SCHUBERT, 2002; JR; MOHAN, 1990).

Embora a doença possa causar a debilitação das árvores, bem como a perda de qualidade e produtividade de frutos, o seu maior impacto está relacionado à perdas econômicas (BOCK; PARKER; GOTTWALD, 2005). De acordo com Jr e Mohan (1990) e Gottwald, Graham e Schubert (2002) uma das medidas que foi tomada para o combate da doença consiste na destruição das árvores contaminadas, cortando e queimando-as.

A doença é causada pela bactéria *Xanthomonas citri subsp. citri* e os sintomas se manifestam por meio de lesões necróticas, salientes, de coloração marrom, muitas vezes circundadas por halo amarelo, em folhas e frutos. Os sintomas são sempre muito característicos, mas podem variar de acordo com o órgão afetado, idade e genótipo do hospedeiro (BITANCOURT, 1957; GOTTWALD; GRAHAM; SCHUBERT, 2002; KOIZUMI, 1985). A Figura (2.1) realça os danos que a doença pode causar nos frutos e folhas das plantas.

Figura 2.1 – Doença cancro cítrico em folhas e frutos de laranjeiras.

O agente causador da doença é uma bactéria Gram-negativa, aeróbia, baciliforme, com um flagelo polar e facilmente isolada e cultivada em laboratório. Esse cultivo pode ser feito em meios de cultura sólidos ou líquidos simples. A faixa ideal de temperatura para tal crescimento é de $28\,^{\circ}\mathrm{C}$ a $32\,^{\circ}\mathrm{C}$. Nessas condições as colônias são visíveis a olho nu após 48 horas de cultivo.

Para o primeiro experimento tratado neste trabalho foram utilizados 16 genótipos de citros, onde 15 deles são comumente chamados de Laranja Doce (citrus sinenses), sendo eles, Pera 460, Rubi 323, Natal 245, Pera 329, Natal M9-324, Pera 331, Westin 340, Westin 16-319, Pera 436, Bahia 25-462, Natal M9-350, Natal 261, Rubi 251, Valência 326, Natal 308, e 1 como Tangerina (citrus reticulata x citrus sinenses), sendo o genótipo Morcot 280.

O experimento em questão foi conduzido pela pesquisadora do Programa de Pós Graduação em Agronomia (PGA-UEM) Andressa Cazetta que fez o levantamento dos dados a partir do método de folhas destacadas e o implantou no Núcleo de Pesquisa em Biotecnologia Aplicada (NBA).

Em cada um dos 16 genótipos foram coletados ramos sadios e com o mesmo estágio de maturação. Esses ramos passaram por lavagem, desinfecção e foi necessária a utilização de uma solução fungicida para evitar a presença de fungos que pudessem diminuir a longevidade das folhas utilizadas no experimento.

Os ramos foram cortados e a inoculação da bactéria *Xanthomonas citri subsp. citri* nas folhas destacadas foi realizada usando a estirpe Xcc 306 e armazenada no NBA-UEM, mantida em geladeira em tampão fosfato alcalino. Visando a reativação da bactéria, a mesma foi semeada em placas de Petri contendo meio Manitol Glutamate Yeast. As placas de Petri foram acondicionadas em estufa bacteriológica por 72 horas a $28\,^{\circ}\mathrm{C}$.

A inoculação da bactéria foi feita utilizando uma agulha esterilizada e as folhas (5 folhas por genótipo) foram perfuradas oito vezes (4 furos em cada lado do limbo foliar). As folhas inoculadas foram mantidas em Tubo Falcon (GREINER) com água da torneira suficiente

para cobrir parte do ramo e metade do pecíolo sem atingir o limbo foliar. O tubo não foi fechado completamente para que a folha continuasse normalmente sua transpiração e, quando necessário, a água de cada tubo foi reposta. Para o armazenamento, os tubos foram mantidos em temperatura ambiente.

Avaliou-se o diâmetro das lesões ocasionadas pela bactéria com o auxílio de um paquímetro eletrônico nos períodos de 3, 7 e 14 dias após a inoculação (DAI). O número total de medidas na amostra foi de 1920, ou seja:

- 16 genótipos de citros;
- 5 folhas destacadas em cada um dos genótipos;
- 8 perfurações em cada uma das folhas destacadas;
- 3 avaliações (em cada um dos dias) em cada perfuração.

Com relação ao segundo experimento foram utilizados 6 genótipos de citros de laranja doce, sendo eles Pera IAC, Irradiada, Pera Ori, Prec Ori, Hamlin e Valência.

O experimento foi conduzido pela pesquisadora Juliana Glória Franco do Programa de Pós Graduação em Agronomia (PGA-UEM), que fez o levantamento dos dados por meio do método de folhas destacadas, onde este também foi implantado no Núcleo de Pesquisa em Biotecnologia Aplicada (NBA).

A coleta das folhas em cada um dos genótipos foi realizada como no primeiro experimento, diferindo-se apenas nas quantidades. Neste caso, em cada um dos 6 genótipos avaliados foram coletadas 8 folhas, onde em cada uma delas foram feitas 3 perfurações que foram avaliadas em 3 tempos diferentes.

As folhas foram inoculadas com auxílio de agulha, em que avaliou-se o diâmetro das lesões causadas pela bactéria com o auxílio de um paquímetro eletrônico nos períodos de 7, 14 e 21 dias após a inoculação da bactéria (DAI). Assim, o número total da amostra foi de 432, ou seja:

- 6 genótipos de citros;
- 8 folhas destacadas em cada um dos genótipos;
- 3 perfurações em cada uma das folhas destacadas;
- 3 avaliações (em cada um dos DAI) em cada perfuração.

2.2 Métodos

2.2.1 Dados Longitudinais e Medidas Repetidas

A estrutura básica que compõe a representação de dados longitudinais com medidas repetidas em g subpopulações, N repetições em cada subpopulação e s ocasiões de avaliação é dada por:

Subpopulações	Unidade Experimental	Ocasiões de Avaliação			
		1	2	$\dots$	s
1	1	y_{111}	y_{121}	$\cdots$	y_{1s1}
1	2	y_{112}	y_{122}	$\cdots$	y_{1s2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1	N_1	y_{11N_1}	y_{12N_1}	$\cdots$	y_{1sN_1}
2	1	y_{211}	221	$\cdots$	y_{2s1}
2	2	y_{212}	y_{222}	$\cdots$	y_{2s2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2	N_2	y_{21N_2}	y_{22N_2}	$\cdots$	y_{2sN_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g	1	y_{g11}	y_{g21}	$\cdots$	y_{gs1}
g	2	y_{g12}	y_{g22}	$\cdots$	y_{gs2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g	N_g	y_{g1N_g}	y_{g2N_g}	$\cdots$	y_{gsN_g}

Para modelar dados com essa característica longitudinal e de medidas repetidas que apresentam possível dependência intra-indivíduos, Cnaan, Laird e Slasor (1997), apresentam um modelo que expande o modelo de regressão linear clássico, permitindo incorporar a falta de independência entre as observações e modelar mais de um termo de erro. Esse modelo é chamado de modelo linear misto (LMM).

O LMM é caracterizado como a 'união' de um modelo linear de efeitos fixos com um modelo linear de efeitos aleatórios. Possuindo dois componentes, um intra-indivíduo e outro entre indivíduos, é utilizado em dados longitudinais e permite que os coeficientes de regressão variem entre os indivíduos (FAUSTO et al., 2008).

Esse modelo assume que o padrão de crescimento tem a mesma forma funcional para todos os indivíduos, porém os indivíduos podem apresentar comportamento longitudinal diferente. Neste sentido, cada um possui sua própria curva de crescimento especificada pelos coeficientes de regressão .

2.2.2 Modelo Linear Misto

Os modelos lineares, nos parâmetros, possuem pelo menos um efeito aleatório causado pelo erro experimental. No caso em que o modelo apresenta outros efeitos aleatórios, além do descrito e em comum com outros efeitos fixos, é denominado modelo linear misto.

Este modelo, pode ser então definido como a união de um modelo de efeitos fixos com um modelo de efeitos aleatórios. Esses efeitos aleatórios devem ser considerados como uma amostra de alguma distribuição populacional convenientemente definida. Uma das características positivas dos modelos mistos é que ao modelar ambos os efeitos, este fornece a possibilidade de modelar também as variâncias e covariâncias e não apenas a média.

É comumente utilizado para a descrição de dados com medidas repetidas, em que as observações são dependentes, flexibilizando assim os modelos com erros correlacionados. A análise de modelos mistos tem como objetivos estimar os parâmetros de covariância, testar hipóteses sobre os parâmetros ou funções dos parâmetros, calcular preditores dos efeitos aleatórios, bem como comparar médias de tratamentos.

O modelo descrito possui várias características positivas. Dentre elas, Laird e Ware (1982) citam o fato de que os dados não precisam ser equilibrados, permite modelagem e análise explícita de variações entre e intra indivíduo, em que os parâmetros individuais, muitas vezes, têm uma interpretação natural relevante para os objetivos do estudo, e suas estimativas podem ser usadas para análises exploratórias .

A literatura sobre essa metodologia também é vasta. Inicialmente, foi estudada por Henderson (1949), mas tornou-se de fato conhecida com o advento das técnicas computacionais mais robustas, pois até então só se fazia uso, em sua maioria, de modelos de efeitos fixos.

Autores como Harville e Mee (1984) propuseram um procedimento de modelo misto para analisar dados categóricos ordenados a fim de prever o valor de uma resposta categórica ordenada a partir do conhecimento de várias variáveis preditivas. Este procedimento se assemelha ao melhor procedimento imparcial linear de Henderson (1975) para prever o valor de uma resposta quantitativa. Os resultados são ilustrados por uma aplicação ao problema de prever o grau de dificuldade que será experimentado por uma vaca leiteira no nascimento de seu bezerro.

Verbeke e Lesaffre (1996), investigaram por meio de dois exemplos práticos o impacto da suposição de normalidade para efeitos aleatórios em suas estimativas no modelo linear de efeitos mistos. Os autores mostraram que se a distribuição de efeitos aleatórios é uma mistura finita de distribuições normais, esses efeitos podem ser mal estimados se a nor-

malidade for assumida, dado que os métodos atuais para inspecionar a adequação das premissas do modelo não são sólidos.

Ferreira e Moraies (2013), avaliaram por meio de um modelo linear misto a influência do uso do café, da espécie *Coffea arabica* no controle de peso de ratos submetidos a diferentes dietas alimentares com e sem extrato aquoso de café. Oliveira et al. (2004), utilizam o modelo linear misto para estudar a variabilidade genética, estimar parâmetros genéticos e realizar a predição de valores genéticos dos indivíduos para fins de seleção, utilizando a metodologia REML/BLUP a partir da avaliação de procedências e progênies de umbuzeiro.

Seja $\mathbf{Y_{ij}}$ a variável resposta do indivíduo i, no instante t_{ij} e $\mathbf{Y_i} = (Y_{i1}, ..., Y_{in_i})^T$ o vetor que contém as n_i observações repetidas do indivíduo i, com $i = 1, ..., N$ e $j = 1, ..., n_i$. Laird e Ware (1982), definiram o modelo linear normal misto como:

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u_i} + \boldsymbol{\varepsilon_i}, \quad i = 1, ..., N, \tag{2.1}$$

em que:

- $\mathbf{Y_i}$: é o vetor $(n_i$ x $1)$ de observações referente a variável resposta da i-ésima unidade experimental;
- $\mathbf{X_i}$: é a matriz $(n_i$ x $p)$ contendo as variáveis explicativas de efeitos fixos;
- $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^T$: é o vetor $(p$ x $1)$ de parâmetros fixos desconhecido;
- $\mathbf{Z_i}$: é a matriz $(n_i$ x $q)$ de variáveis explicativas contendo os efeitos aleatórios;
- $\mathbf{u_i} = (u_{i1}, ..., u_{iq})^T$: é o vetor $(q$ x $1)$ de efeitos aleatórios desconhecido;
- $\boldsymbol{\varepsilon}_i$: é o vetor $(n_i$ x $1)$ de erros residuais.

Marginalmente, $\mathbf{u_i} \sim N_q(0, \mathbf{D})$ e $\varepsilon_i \sim N_{n_i}(0, \mathbf{R_i})$, sendo $\mathbf{R_i}$ uma matriz de covariâncias de dimensão $(n_i$ x $n_i)$, que depende de i por meio da sua dimensão n_i, mas o conjunto de parâmetros desconhecidos em $\mathbf{R_i}$ não dependerá de i e $\mathbf{D}$ a matriz de covariâncias de dimensão $(q$ x $q)$. Ambas simétricas e geralmente inversíveis. É importante destacar também que $\mathbf{u_i}$ e ε_i são mutuamente independentes.

Em um contexto de medidas repetidas, as variâncias dos efeitos aleatórios $\mathbf{u_i}$ medem a variabilidade das trajetórias longitudinais entre os indivíduos e as variâncias dos erros residuais medem a variabilidade das observações repetidas intra-indivíduos (MANCO, 2013).

O modelo linear misto, é comumente especificado em termos das respostas $\mathbf{Y_i}$ condicionadas aos efeitos aleatórios $\mathbf{u_i}$ (BARBOSA, 2009). Se $\mathbf{Y_i}$ é um vetor de medidas repetidas para o $i-$ésimo indivíduo, então:

$$\mathbf{Y_i}|\mathbf{u_i} \sim N(\mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u_i}, \mathbf{R_i}), \qquad \text{e} \qquad \mathbf{u_i} \sim N(0, \mathbf{D}).$$

Apesar disso, as inferências são realizadas por meio do modelo marginal, sendo que $\mathbf{Y_i}$ segue distribuição normal, com média $\mathbf{X_i}\alpha$ (que pode ser definida como a média sobre todos os efeitos aleatórios) e variância $\mathbf{R_i} + \mathbf{Z_i}\mathbf{D}\mathbf{Z_i^T}$, ou seja,

$$\mathbf{Y_i} \sim N(\mathbf{X_i}\boldsymbol{\alpha}, \mathbf{R_i} + \mathbf{Z_i}\mathbf{D}\mathbf{Z_i^T}).$$

É fácil ver que marginalmente, os efeitos fixos estão presentes apenas na média, enquanto os aleatórios estão presentes na variância.

Observa-se que um bom ajuste do modelo está diretamente ligado à escolha das estruturas de variâncias e covariâncias das variáveis aleatórias, pois a covariância entre as observações obtidas na mesma unidade experimental pode ser modelada indiretamente por meio dos efeitos aleatórios $(\mathbf{u_i})$.

A forma matricial para as estruturas de variâncias e covariâncias desse modelo nas observações repetidas do indivíduo i, é dada por:

$$Cov(\mathbf{Y_i}) = \mathbf{V_i} = \mathbf{R_i} + \mathbf{Z_i}\mathbf{D}\mathbf{Z_i^T}.$$

Laird e Ware (1982), citam que um modelo mais simples pode surgir quando $\mathbf{R_i} = \sigma^2 I$ (I sendo a matriz identidade), e este é chamado de modelo de independência condicional, pois implica que as n_i respostas no indivíduo i são independentes, condicionais a $\mathbf{u_i}$ e α.

Existem alguns tipos de estruturas de matrizes de covariâncias que podem ser utilizadas dependendo do contexto do problema, da estrutura dos dados e algumas vezes até da disponibilidade computacional. As mais utilizadas são as não estruturadas, de simetria composta, autorregressivas e toeplitz e estas são descritas brevemente no ANEXO A. A modificação dessas estruturas possibilita a inclusão de correlações entre as observações.

Os parâmetros do modelo a serem estimados são os parâmetros de efeitos fixos α, as variâncias e covariâncias da matriz $\mathbf{D}$ de efeitos aleatórios, bem como as variâncias e covariâncias da matriz $\mathbf{R}$ de erros residuais. Essas estimações podem ser realizadas por vários métodos. Searle, Casella e McCulloch (2009) e Patterson e Thompson (1971) citam em seus trabalhos os métodos de máxima verossimilhança e máxima verossimilhança restrita.

Ao utilizar o método da máxima verossimilhança, estima-se a variância e assume-se que não existe erro na estimativa da média. Em contrapartida, o método da máxima verossimilhança restrita produz estimativas não viciadas da variância, removendo-se o vício que existe na estimação da média (FAUSTO et al., 2008).

Ainda dentro da abordagem de modelos mistos, encontra-se os modelos lineares generalizados mistos (GLMM), que são uma extensão dos LMM. Estes, avaliam a variável de interesse mesmo que esta não tenha um comportamento de uma variável normalmente distribuída. A única exigência é de que a distribuição adotada para essa variável pertença a família exponencial de distribuições. Tal extensão, permite expandir a quantidade de análises e inferências já feitas nos LMM, incluindo dados com comportamento assimétrico.

2.2.3 Modelo Linear Generalizado Misto

A classe dos modelos lineares generalizados (GLM) introduzida por Nelder e Wedderburn (1972), considera apenas o estudo de variáveis com efeitos fixos. Uma classe de modelos que permite os efeitos aleatórios (juntamente com os fixos) no preditor linear (AGRESTI, 2018) é uma extensão natural a se pensar. Ao incorporar os efeitos aleatórios no preditor linear é possível modelar a estrutura de correlação entre as observações que pertencem ao mesmo indivíduo.

Dessa forma, os GLMM, assim como os LMM, buscam descrever o comportamento de uma variável aleatória, por meio de variáveis explicativas, mas se diferem quanto a distribuição adotada para tal variável. Enquanto nos LMM a variável aleatória deve estar associada a distribuição normal, nos GLMM essa exigência não é necessária, ou seja, a variável aleatória pode assumir qualquer distribuição, desde que essa pertença a família exponencial.

Esse tipo de abordagem foi proposta por Breslow e Clayton (1993) e foi ganhando espaço em outros estudos, como os de Fong, Rue e Wakefield (2010) e Zhao et al. (2006), que introduziram conceitos de inferência Bayesiana aos GLMM, Lee et al. (2012) que propuseram um GLMM para dados binários longitudinais, Michel, Brun e Makowski (2017) introduziram uma estrutura baseada em GLMM para analisar pesquisas de pragas e doenças. Agresti (2018), em um dos capítulos de seu trabalho, exemplifica um GLMM logístico para dados binários combinados, assim como faz uso de extensões para GLMM que podem possuir mais de um termo de efeito aleatório no modelo, entre outros.

Para a especificação do GLMM, considere a distribuição condicional de $\mathbf{Y}$ dado $\mathbf{u}$, sendo que $\mathbf{Y}$ é o vetor de respostas assumindo consistir de elementos condicionalmente independentes e com função densidade pertencente a família exponencial. Nesse cenário,

tem-se que:

$$\mathbf{Y_i}|\mathbf{u} \sim \quad família \;\; exp(y_i;\theta;\phi)$$

$$f_{\mathbf{Y_{i|u}}}(y_i|\mathbf{u}) = \exp\left\{\frac{w_i}{\phi}[y_i\theta_i - b(\theta_i)] + c(y_i;\phi)\right\}, \tag{2.2}$$

em que $f_{\mathbf{Y_{i|u}}}(y_i|\mathbf{u})$ é a função densidade (ou probabilidade) de $\mathbf{Y}_i$ dado u da família exponencial de distribuições , com θ_i sendo o parâmetro canônico, w_i uma constante conhecida e ϕ é um parâmetro de dispersão ou de escala.

Os GLMM inclui ao modelo um preditor linear η_i, que é utilizado para modelar a relação entre a resposta e os efeitos fixos e aleatórios, uma função de ligação $g(.)$ conhecida, que modela a relação entre o preditor e a média condicional μ_i, bem como uma função de variância para modelar a variabilidade residual.

Dessa forma,

$$\mathbb{E}(\mathbf{Y_i}|\mathbf{u}) = \mu_i = b'(\theta_i)$$

e

$$g(\mu_i) = \eta_i = \mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u}$$

em que

- $\mathbf{X_i}$: Matriz de variáveis explicativas de efeitos fixos;
- $\boldsymbol{\alpha}$: é o vetor de parâmetros fixos;
- $\mathbf{Z_i}$: Matriz de variáveis explicativas de efeitos aleatórios;
- $\mathbf{u}$: é o vetor de efeitos aleatórios.

Ao considerar o modelo condicional em 2.2, defini-se a média e a variância, respectivamente como:

$$\mathbb{E}(\mathbf{Y_i}) = \mathbb{E}(\mathbb{E}(\mathbf{Y_i}|\mathbf{u})) = \mathbb{E}(\mu_i) = \mathbb{E}(g^{-1}(\mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u})).$$

e

$$\begin{aligned} V(\mathbf{Y_i}) &= V[\mathbb{E}(\mathbf{Y_i}|\mathbf{u})] + \mathbb{E}[V(\mathbf{Y_i}|\mathbf{u})] \\ &= V[\mu_i] + \mathbb{E}[a_i(\phi)V(\mu_i)] \\ &= V[g^{-1}(\mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u})] + \mathbb{E}\{a_i(\phi)V[g^{-1}(\mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u})]\}, \quad a_i = \frac{w_i}{\phi}. \end{aligned}$$

Ao incorporar efeitos aleatórios ao modelo, inclui-se a correlação entre as observações que possuam algum efeito em comum. Logo, a covariância e a correlação são definidas respectivamente da seguinte forma:

$$\begin{aligned} Cov(\mathbf{Y_i}, \mathbf{Y_j}) &= Cov[\mathbb{E}(\mathbf{Y_i}|\mathbf{u}), \mathbb{E}(\mathbf{Y_j}|\mathbf{u})] + \mathbb{E}[Cov(\mathbf{Y_i}, \mathbf{Y_j}|\mathbf{u})] \\ &= Cov(\mu_i, \mu_j) + \mathbb{E}(0) \\ &= Cov[g^{-1}(\mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u}), g^{-1}(\mathbf{X_j}\boldsymbol{\alpha} + \mathbf{Z_j}\mathbf{u})]. \end{aligned}$$

e

$$Corr(\mathbf{Y_i}, \mathbf{Y_j}) = \frac{Cov(g^{-1}(\mathbf{X_i}\boldsymbol{\alpha} + \mathbf{Z_i}\mathbf{u}), g^{-1}(\mathbf{X_j}\boldsymbol{\alpha} + \mathbf{Z_j}\mathbf{u}))}{\sqrt{Var(\mathbf{Y_i})}\sqrt{Var(\mathbf{Y_j})}}.$$

Sabendo que é possível em um único modelo termos ambos os efeitos aleatórios, ou seja, modelar as estruturas de covariâncias $\mathbf{D}$ e $\mathbf{R}$, pode-se definir a estrutura de covariância de $\mathbf{Y}$ para este caso, onde esta é dada por $Cov(\mathbf{Y}) = \mathbf{ZDZ^T} + \mathbf{R}$.

Assim como nos LMM, nos GLMM a escolha das estruturas de variâncias e covariâncias das variáveis aleatórias é de suma importância para uma modelagem satisfatória. São inúmeras as opções para esse tipo de estudo listadas na literatura. De acordo com Xavier (2000) tem-se algumas opções dessas estruturas, como por exemplo, não estruturada, auto-regressiva de primeira ordem, auto-regressiva de primeira ordem heterogênea, auto-regressiva de primeira ordem médias móveis, simetria composta e toeplitz. Estas são expostas de forma mais detalhadas no ANEXO A.

Apesar de algumas matrizes serem mais utilizadas para alguns tipos específicos de dados, todas as matrizes citadas podem ser aplicadas em GLMM. Gbur et al. (2012) cita que a escolha da matriz de covariâncias mais adequada pode ser feita através dos critérios de informação como AIC (AKAIKE, 1974) ou BIC (SCHWARZ et al., 1978).

Estes, podem ser usados para comparar estruturas de covariância, desde que a parte dos efeitos fixos do modelo seja a mesma para todas as estruturas em consideração. Os critérios de informação são calculados para cada modelo candidato e seus valores são comparados . O modelo com estrutura de covariâncias que apresente menor valor de AIC ou BIC é considerado o mais adequado.

O Teste da Razão de Verossimilhança (TRV), também pode ser usado na escolha da matriz de covariâncias mais adequada. Este, permite comparar modelos dois a dois, em que um deles é a versão restrita do outro, ou seja, modelo completo, obtido da maximização

em todo o espaço de parâmetros e modelo reduzido, obtido da maximização no limite de restrição.

Outro aspecto que um GLMM tem em comum com um LMM está nos parâmetros a serem estimados. Estima-se os parâmetros de efeitos fixos, as variâncias e covariâncias da matriz **D** de efeitos aleatórios, bem como as variâncias e covariâncias da matriz **R** de erros residuais. Entretanto, os métodos utilizados para tal estimação se diferem.

Enquanto no LMM utiliza-se a estimação por máxima verossimilhança ou máxima verossimilhança restrita, nos GLMM esses métodos não podem ser utilizados pois geralmente a função de verossimilhança não possui forma fechada, o que gera integrais que não podem ser resolvidas analiticamente. Esse problema pode ser resolvido através da maximização da função de verossimilhança penalizada ou por meio da h-verossimilhança (verossimilhança hierárquica) que são consideradas praticamente equivalentes do ponto de vista de maximização (RIGBY; STASINOPOULOS, 2005).

2.2.4 Modelo Aditivo Generalizado para Localização, Escala e Forma

Uma extensão do GLMM é o Modelo Aditivo Generalizado para Locação, Escala e Forma (GAMLSS). Um dos principais ganhos ao se utilizar esse tipo de modelo é que a suposição da variável resposta ter de seguir distribuição que pertença a família exponencial é relaxada e substituída por uma família de distribuições mais geral.

Assim, qualquer distribuição de probabilidade implementada no *Software R* por meio do pacote `gamlss` (STASINOPOULOS; RIGBY et al., 2007) pode ser utilizada na modelagem, incluindo aquelas com altos níveis de assimetria e curtose, no caso contínuo e também discreto, permitindo assim uma maior flexibilidade no ajuste de modelos.

De acordo com Stasinopoulos, Rigby et al. (2007) os GAMLSS são modelos de regressão semiparamétricos. São paramétricos, no sentido de que requerem uma suposição de distribuição paramétrica para a variável resposta e "semi" no sentido de que a estimação dos parâmetros da distribuição, como funções de variáveis explicativas, pode envolver o uso de funções de suavização não-paramétricas. Foi introduzido por Rigby e Stasinopoulos (2001, 2005) e Akantziliotou e Stasinopoulos (2002), com o objetivo de superar algumas das limitações que eram associadas aos GLM e aos Modelos Aditivos Generalizados (GAM).

Além dos pontos já citados, o GAMLSS assume observações independentes $\mathbf{Y}_i$, para $i = 1, .., n$ com função densidade de probabilidade $f(\mathbf{Y}_i|\theta^i)$ condicional com

$$\theta^{\mathbf{i}} = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i),$$

em que $\theta^{\mathbf{i}}$ é um vetor de quatro parâmetros e cada um pode ser uma função das variáveis explicativas. Os primeiros dois parâmetros μ_i e σ_i são geralmente caracterizados como parâmetros de localização e escala da distribuição da população, enquanto os parâmetros restantes, se houver, são caracterizados como parâmetros de forma, por exemplo, parâmetros de assimetria e curtose.

Dessa forma, além de estimar a média em função das variáveis explicativas e efeitos aleatórios, é possível expandir o componente sistemático do modelo e estimar outros parâmetros da distribuição de $\mathbf{Y}_i$. Além disso, os preditores são formados por funções lineares e/ou não lineares, paramétricas e/ou aditivas não-paramétricas. A estimação por máxima verossimilhança penalizada é utilizada para ajustar os modelos, onde os algoritmos utilizados nessa estimação serão detalhados de forma sucinta na Equação (2.3) .

Seja $\mathbf{Y_i}^T = (Y_{i1}, ..., Y_{in_i})$, em que n é o comprimento do vetor da variável resposta. Também para $k = 1, 2, 3, 4$, sejam $g_k(.)$ funções de ligação monótonas conhecidas, com o $k-$ésimo parâmetro θ_k relacionado com as variáveis explicativas dos modelos aditivos semi-paramétricos, definidas da seguinte forma:

$$\begin{aligned}
g_1(\mu) &= \eta_1 = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \\
g_2(\sigma) &= \eta_2 = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}) \\
g_3(\nu) &= \eta_3 = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}) \\
g_4(\tau) &= \eta_4 = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}),
\end{aligned} \tag{2.3}$$

em que $\mu, \sigma, \nu, \tau, \eta_\neg$ e $\mathbf{x}_{jk}$ são vetores de comprimento n, com $j = 1, ..., J_k$ e $k = 1, ..., 4$. A função h_{jk} é aditiva não-paramétrica da variável explicativa $\mathbf{X}_{jk}$, avaliada em $\mathbf{x}_{jk}$. Os vetores explicativos $\mathbf{x}_{jk}$ são assumidos como fixos e conhecidos, assim como $\mathbf{X}_k$ são fixos na matriz de planejamento, enquanto $\beta_{\mathbf{k}}$ são os vetores de parâmetros de regressão. O modelo (2.3) é chamado de modelo aditivo generalizado para locação, escala e forma e pode ser estendido a fim de permitir que efeitos aleatórios sejam incluídos. Para mais detalhes, ver Rigby e Stasinopoulos (2005).

Dentre todas as distribuições implementadas no pacote `gamlss` do *Software R* e pertencentes a classe dos GAMLSS, há três tipos diferentes de famílias de distribuições: contínuas, discretas e a mistura de distribuições. A utilização destas, depende do tipo de problema que o pesquisador pretende resolver, da natureza dos dados, assim como do comportamento da variável resposta que está sendo modelada. Detalhes referentes a

nomenclaturas utilizadas no *R*, espaço paramétrico, domínio das funções, funções de ligação e demais informações sobre cada uma das distribuições implementadas, podem ser encontrados no trabalho de Rigby et al. (2019).

Por todos os pontos positivos já mencionados, optou-se neste trabalho por fazer todos os ajustes por meio do pacote `gamlss` por ser mais completo e possuir todas as distribuições de interesse já implementadas.

2.2.4.1 Estimação dos Parâmetros do Modelo

O pacote `gamlss`, do *Software R* permite modelar uma família de distribuições, em geral com no máximo quatro parâmetros, sendo estes os parâmetros de locação, escala e forma (assimetria e curtose). As distribuições implementadas podem ser encontradas por meio da função `gamlss.family`. Os algoritmos atuais para estimação dos parâmetros de um modelo utilizados pelo `gamlss` são RS, CG e MISTO. O tipo de método a ser usado fica a critério do pesquisador e pode ser alterado usando o argumento `method`.

O método RS refere-se ao algoritmo de Rigby e Stasinopoulos (1996) utilizado para ajustar modelos aditivos de média e dispersão (e não usa derivadas cruzadas). Esse método é o padrão utilizado e não requer valores iniciais precisos para os parâmetros para garantir a convergência.

O método CG refere-se ao algoritmo de Cole e Green (1992), este utiliza a primeira e (esperada ou aproximada) segunda derivada cruzada da função de verossimilhança em relação aos parâmetros . O método CG tende a ser melhor para distribuições com estimativas de parâmetros que possam ser altamente correlacionados.

Por fim, o método MISTO é uma mistura dos métodos RS e CG, onde utiliza o algoritmo RS duas vezes antes de mudar para o algoritmo CG por até 10 iterações extras. É empregado para a distribuição adotada, fazendo com que a convergência seja acelerada.

Portanto, para os ajustes dos modelos das duas aplicações, utilizou-se o pacote `gamlss` do Software *R*, versão 3.6.3. A estimação dos parâmetros foi feita por meio da função `gamlss` deste mesmo pacote e o método utilizado em todas as análises foi o RS de Rigby e Stasinopoulos (1996a). Mais detalhes do método serão dados na sequência.

2.2.4.2 O Método RS

Com o intuito de apresentar maiores informações sobre o processo de estimação utilizado neste trabalho, é exemplificado a seguir o método RS, seguindo o processo detalhado em Rigby e Stasinopoulos (2005).

O algoritmo em questão possui um ciclo externo que maximiza a verossimilhança penalizada em relação a β_k e γ_{jk}, para $j = 1, ..., J_k$, sucessivamente para cada θ_k, com $k = 1, ..., p$. Em cada um dos cálculos do algoritmo, são usados todos os valores atuais atualizados de todas as quantidades.

Vale ressaltar, que o algoritmo RS não é um caso particular do algoritmo CG, ao passo que no RS a matriz de peso diagonal $\mathbf{W}_{kk}$ é atualizada dentro do ajuste de cada parâmetro θ_k, enquanto no algoritmo CG todas as matrizes de peso $\mathbf{W}_{ks}$ para $k = 1, ..., p$ e $s = 1, ..., p$ são avaliadas após o ajuste de todos os θ_k.

Os passos utilizados na estimação pelo algoritmo RS de acordo com os autores supracitados são:

- **Etapa 1:** Inicializar o ajuste dos valores $\theta_k^{(1,1)}$ e efeitos aleatórios $\gamma_{jk}^{(1,1,1)}$ para $j = 1, ..., J_k$ e $k = 1, ..., p$. Avaliar então, os preditores lineares iniciais $\eta_k^{(1,1)} = g_k(\theta_k^{(1,1)})$ para $k = 1, ..., p$.

- **Etapa 2:** Iniciar o ciclo externo $r = 1, 2, ...$ até a convergência.

  - **(a)** Iniciar o ciclo interno $i = 1, 2, ...$ até a convergência.

    (i) Avaliar o atual $\mathbf{u}_k^{(r,i)}$, $\mathbf{W}_{kk}^{(r,i)}$ e $\mathbf{z}_k^{(r,i)}$;

    (ii) Inciar o ciclo de Backfitting[1] $m = 1, ...$ até a convergência;

    (iii) Regredir os resíduos parciais atuais $\epsilon_{0k}^{(r,i,m)} = \mathbf{z}_k^{(r,i)} - \sum_{j=1}^{J_k} \cdot \mathbf{Z}_{jk}\gamma_{jk}^{(r,i,m)}$ contra a matriz $\mathbf{X}_k$, usando os pesos iterativos $\mathbf{W}_{kk}^{(r,i)}$ para obter as estimativas dos parâmetros $\beta_k^{(r,i,m+1)}$;

    (iv) Para $j = 1, ..., J_k$ analisar os resíduos parciais $\epsilon_{jk}^{(r,i,m)} = \mathbf{z}_k^{(r,i)} - \mathbf{X}_k\beta_k^{(r,i,m+1)} - \sum_{j=1,t\neq j}^{J_k} \mathbf{Z}_{tk}\gamma_{tk}^{(r,i,c)}$ para obter a matriz de encolhimento (suavização) $\mathbf{S}_{jk}$ que é dada por:

    $$\mathbf{S}_{jk} = \mathbf{Z}_{jk}(\mathbf{Z}_{jk}^T\mathbf{W}_{kk}\mathbf{Z}_{jk} + \mathbf{G}_{jk})^{-1}\mathbf{Z}_{jk}^T\mathbf{W}_{kk},$$

    alcançando assim, o termo preditor aditivo atualizado (e atual) $\mathbf{Z}_{jk}\gamma_{jk}^{(r,i,m+1)}$;

    (v) Fim do ciclo de Backfitting na convergência de $\beta_k^{(r,i,...)}$ e $\mathbf{Z}_{jk}\gamma_{jk}^{(r,i,...)}$ e definindo $\beta_k^{(r,i+1)} = \beta_k^{(r,i,...)}$ e $\gamma_{jk}^{(r,i+1)} = \gamma_{jk}^{(r,i,...)}$ para $j = 1, ..., J_k$ e de outra forma atualizar m e continuar o ciclo de Backfitting;

[1] Backfitting: Procedimento iterativo para ajuste de modelos aditivos em que, cada passo, um componente é estimado, mantendo os outros componentes fixos, iterando até a convergência. É utilizado, pois se torna mais fácil estender o algoritmo para que novos termos aditivos possam ser incluídos. Mais informações podem ser vistas em Stasinopoulos, Rigby et al. (2007) e um exemplo de prova geométrica simples, porém geral da convergência do Backfitting para alguns casos podem ser encontrados em Ansley e Kohn (1994).

(vi) Calcular o $\eta_k^{(r,i+1)}$ e $\theta_k^{(r,i+1)}$.

– **(b)** Terminar o ciclo interno na convergência de $\beta_k^{(r,i,...)}$ e os termos preditores aditivos $\mathbf{Z}_{jk}\gamma_{jk}^{(r,...)}$ e definir $\beta_k^{(r+1,1)} = \beta_k^{(r,...)}$, $\gamma_{jk}^{(r+1,1)} = \gamma_{jk}^{(r,...)}$ para $j = 1, ..., J_k$, $\eta_k^{(r+1,1)} = \eta_k^{(r,...)}$ e $\theta_k^{(r+1,1)} = \theta_k^{(r,...)}$, caso contrário atualizar i e continuar o ciclo interno.

- **Etapa 3:** Atualizar o valor de k;
- **Etapa 4:** Terminar o ciclo externo se a mudança na verossimilhança penalizada for suficientemente pequena, caso contrário, atualizar r e continuar o ciclo externo.

Mais detalhes sobre o processo de estimação dos parâmetros através do método RS podem ser encontrados nos trabalhos de Stasinopoulos, Rigby et al. (2007) e Rigby e Stasinopoulos (2005).

2.2.4.3 Seleção de Modelos

Muitas vezes faz-se uso de métodos gráficos para selecionar modelos, ou seja, o pesquisador analisa os gráficos de diagnóstico dos modelos candidatos e escolhe o "melhor". Esta análise fica subjetiva ao olhar do pesquisador e nem sempre é suficiente para a escolha de um modelo realmente adequado.

Faz-se necessário então o uso de outros critérios para essa seleção. Um critério que pode ser usado no contexto desse trabalho é o Desvio Global (também chamado de deviance e presente no pacote `gamlss` através da função `deviance`), que é definido como menos duas vezes o logaritmo da função de verossimilhança (RIGBY; STASINOPOULOS, 2005), isto é,

$$GD = -2l(\hat{\theta}).$$

Uma segunda opção para a seleção de modelos é o Critério de informação de Akaike generalizado (GAIC), que leva em consideração o número de parâmetros do modelo, bem como os graus de liberdade utilizados no ajuste. Rigby e Stasinopoulos (2005), definem o GAIC como sendo uma estatística obtida pela adição de GD de uma penalidade fixada p para cada grau de liberdade (gl) usado no modelo. Logo:

$$GAIC(p) = GD + p \text{ x } gl.$$

O AIC e o BIC são casos especiais do $GAIC(p)$, quando $p = 2$ e $p = log(n)$, respectivamente (RIGBY; STASINOPOULOS, 2005). Esses critérios são comumente utilizados pois ponderam parcimônia (menor número de parâmetros) com melhor adequação (menores

desvios). A comparação é feita por meio dos valores obtidos nas estatísticas em cada modelo testado, onde o modelo que resultar em um menor valor é tratado como sendo o de melhor ajuste (PAIVA; FREIRE; CECATTI, 2008).

Assim, a seleção de modelos em ambas aplicações foram feitas através dos critérios AIC, BIC e GD, já mencionados. A partir da análise desses valores, a parcimônia foi que determinou o modelo final escolhido para a modelagem, bem como o estudo sobre quais variáveis seriam de efeito fixo ou aleatório dentro da abordagem de modelos mistos.

Vale ressaltar também, que para as cinco distribuições utilizadas na primeira aplicação, foram ajustados os modelos finais com as estruturas de variâncias e covariâncias ARMA(1,1) e SC. Diferentemente da primeira, na segunda aplicação uma maior quantidade de estruturas de variâncias e covariâncias puderam ser testadas ao passo que as avaliações dos diâmetros das lesões eram igualmente espaçadas. Os resultados obtidos não foram significativos quando comparados à não estruturada (UN). Dessa forma, a estrutura UN, que assume correlações independentes e calculadas a partir dos dados foi a utilizada nas duas aplicações.

2.2.4.4 Diagnóstico do Modelo

Há algumas opções de ferramentas estatísticas quando o assunto é a análise de resíduos de um modelo. Segundo Dunn e Smyth (1996), os resíduos quantílicos aleatorizados podem generalizar qualquer um dos métodos de diagnóstico usuais que usam resíduos.

Os autores definem que a distribuição destes resíduos converge para uma distribuição normal padrão à medida que n cresce, para qualquer distribuição de probabilidade para a variável resposta. Dado que $F(y_i; \mu_i, \phi_i)$ é a função distribuição acumulada de $P(\mu, \phi)$. Se F é contínua e $F(y_i; \mu_i, \phi_i)$ é uniformemente distribuída no intervalo unitário, os resíduos são definidos como:

$$\hat{r}_{q,i} = \Phi^{-1} F(y_i; \hat{\mu}_i, \hat{\phi}_i), \tag{2.4}$$

em que, $\Phi(.)$ corresponde à função da distribuição acumulada da distribuição normal padrão e $F(y_i; \mu_i, \phi_i)$ é definido como a função de distribuição acumulada da distribuição da variável resposta y_i para $i = 1, \cdots, n$.

No caso em que a distribuição de F é discreta, os resíduos são definidos da seguinte forma:

$$\hat{r}_{q,i} = \Phi^{-1}\{u_i\}, \tag{2.5}$$

em que, u_i é uma variável aleatória que segue distribuição uniforme no intervalo $[F(y_i - 1|\hat{\mu}_i), F(y_i|\hat{\mu}_i)]$, com $i = 1, \cdots n$.

Para fazer tal análise graficamente, a saída padrão nos modelos ajustados no pacote `gamlss`, através da função `plot` é composta por quatro gráficos. Apresenta os gráficos de resíduos versus valores ajustados, resíduos versus um índice ou covariável, densidade estimada de Kernel dos resíduos e gráfico quantil-quantil normal dos resíduos.

Nesse sentido, um modelo se adequará bem aos dados se os seus resíduos quantílicos aleatorizados apresentarem distribuição aproximadamente normal padrão, ou seja, $N(0,1)$ com coeficiente de assimetria próximo de 0 e curtose próxima de 3. Essas exigências são válidas mesmo quando a distribuição do modelo não é normal. Ao fazer a análise gráfica de resíduos, espera-se que a distribuição destes esteja o mais próximo possível de sua distribuição de referência.

O gráfico denominado Worm Plot (BUUREN; FREDRIKS, 2001) (presente no pacote `gamlss` através da função `wp`) também se torna uma opção, pois permite avaliar em que faixa do intervalo da variável explicativa o modelo não se ajusta adequadamente. Nesse gráfico são apresentados os limites superior e inferior do intervalo de confiança de 95%, delimitando então a região interna onde os pontos devem se localizar para o caso de um bom ajuste. No caso em que uma quantidade maior que 5% dos pontos estiver fora dessa região interna delimitada e consequentemente se distanciando da reta horizontal em torno de zero, pode-se dizer que o modelo utilizado não é adequado.

2.3 Possíveis distribuições utilizadas na análise dos dados

2.3.1 Distribuição Gama

A distribuição Gama (GA) possui algumas parametrizações descritas na literatura. Aqui será apresentada a parametrização dada em Stasinopoulos, Rigby e Akantziliotou (2008). Considerando uma variável aleatória Y que segue distribuição GA, comumente denotada por $GA(\mu, \sigma^2)$ reparametrizada em função de sua média , ou seja, fazendo $\mu = \alpha\beta$ e $\sigma^2 = \frac{1}{\alpha}$, possui função densidade de probabilidade (f.d.p.) expressa por:

$$f(y;\mu,\sigma^2) = \frac{1}{(\mu\sigma^2)^{1/\sigma^2}} \cdot \frac{y^{\frac{1}{\sigma^2}-1} e^{-y/\sigma^2\mu}}{\Gamma(1/\sigma^2)}, \quad \text{em que } \Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx, \forall\alpha > 0, \quad (2.6)$$

para $y > 0$, em que $\mu > 0$ e $\sigma > 0$. Os parâmetros μ e σ são denominados parâmetros de escala e forma da distribuição. Tem-se que o valor esperado de Y é dado por $\mathbb{E}(Y) = \mu$ e respectiva variância sendo $Var(Y) = \mu^2\sigma^2$.

A versatilidade da distribuição GA é indiscutível, uma vez que alterados os valores de seus parâmetros pode se obter outras distribuições oriundas desta. Um exemplo de tal fato

é que a distribuição qui-quadrado com r graus de liberdade pode ser considerada um caso particular da distribuição GA para valores de $\mu = \frac{r}{2}$ e $\sigma^2 = \frac{1}{2}$, ou seja,

$$X \sim \chi^2_r \quad \Leftrightarrow \quad X \sim GA\left(\frac{r}{2}, \frac{1}{2}\right).$$

Fixado o valor de $\mu = 1$ e variando os valores de σ é possível observar na Figura (2.2) diferentes formas para essa distribuição, ficando notório que para valores menores de σ a distribuição começa a mostrar uma certa simetria em torno da média, ou seja, se aproxima assintoticamente de uma distribuição $N(\mu, \sigma^2\mu^2)$.

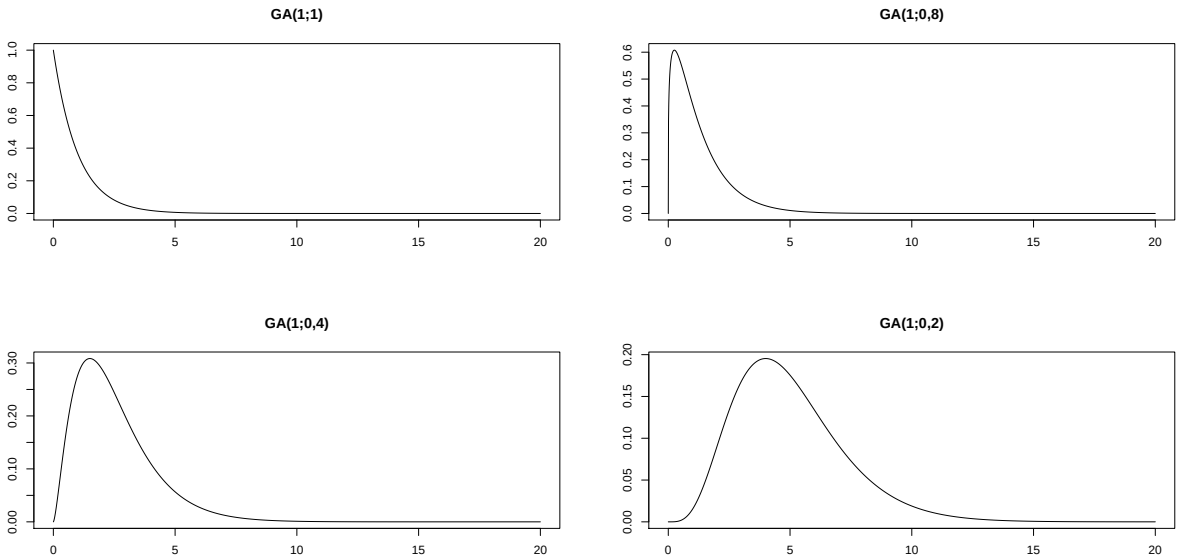

Figura 2.2 – Distribuição Gama para diferentes valores de σ.

Considerando que a distribuição GA pertence a família exponencial de distribuições,esta pode ser vista na seguinte forma:

$$\begin{aligned}
f(y;\mu,\sigma^2) &= \exp\left\{\log\left(\frac{1}{(\mu\sigma^2)^{1/\sigma^2}}\frac{y^{\frac{1}{\sigma^2}-1}e^{-y/\sigma^2\mu}}{\Gamma(1/\sigma^2)}\right)\right\} \\
&= \exp\left\{\log\left(\frac{1}{(\mu\sigma^2)^{(1/\sigma^2)}}\right) + \log(y^{\frac{1}{\sigma^2}-1}e^{-y/\sigma^2\mu}) - \log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right)\right\} \\
&= \exp\left\{\log\left(\frac{1}{\mu\sigma^2}\right)^{-\frac{1}{\sigma^2}} + \log(y^{\frac{1}{\sigma^2}-1}) + \log(e^{-y/\sigma^2\mu}) - \log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right)\right\} \\
&= \exp\left\{-\frac{1}{\sigma^2}\log\left(\frac{1}{\mu\sigma^2}\right) + \left(\frac{1}{\sigma^2}-1\right)\log(y) - \frac{y}{\sigma^2\mu} - \log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right)\right\},
\end{aligned}$$

em que, $\phi = \left(\frac{1}{\sigma^2}\right)^{-1}$, $c(y,\phi) = \left(\frac{1}{\sigma^2} - 1\right)\log(y) - \log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right)$, $\theta = -\frac{1}{\mu\sigma^2}$ e $b(\theta) = -\log(-\theta)$.

Nesse sentido, é possível definir seu respectivo valor esperado e variância, que podem ser obtidos através da própria família exponencial. Assim, a média e a variância de uma variável aleatória Y cuja distribuição pertence à família exponencial, na forma canônica usada por McCullagh e Nelder (1989), são dadas por

$$\mathbb{E}(Y) = b'(\theta) = \mu \tag{2.7}$$

$$Var(Y) = a(\phi)\cdot b''(\theta) = a(\phi)\cdot V(\mu) = \sigma^2 \tag{2.8}$$

Dessa forma, os respectivos valores de média e variância para a distribuição GA, são:

- Média: Dado que $\theta = -\frac{1}{\mu\sigma^2}$ e $b(\theta) = -\log(-\theta)$, tem-se:

$$\begin{aligned}
\mathbb{E}(Y) &= b'(\theta) \\
&= \left[-\log\left(\frac{1}{\mu\sigma^2}\right)\right]' \\
&= -\mu\sigma^2\cdot\left(-\frac{\sigma^2}{\sigma^4}\right) \\
&= \mu.
\end{aligned}$$

- Variância:

$$\begin{aligned}
Var(Y) &= a(\phi)\cdot b''(\theta) \\
&= a(\phi)\cdot V(\mu) \\
&= \left(\frac{1}{\sigma^2}\right)^{-1}\cdot\mu^2 \\
&= \mu^2\sigma^2.
\end{aligned}$$

Com relação a estimação dos parâmetros, são comumente utilizados os métodos de momentos e de máxima verossimilhança. É fácil citar vantagens de um estimador de máxima verossimilhança, dentre elas a sua propriedade de suficiência, invariância e não ser viesado assintoticamente, por exemplo. Portanto, define-se a função de verossimilhança da distribuição GA com parâmetros μ e σ , como:

$$L(y_i;\mu,\sigma^2) = \prod_{i=1}^{n}\left[\frac{1}{(\mu\sigma^2)^{1/\sigma^2}}\cdot\frac{y^{\frac{1}{\sigma^2}-1}e^{-y/\sigma^2\mu}}{\Gamma(1/\sigma^2)}\right]. \tag{2.9}$$

Aplicando-se o logaritmo, tem-se a função log-verossimilhança:

$$\begin{aligned}
l(y_i;\mu,\sigma^2) &= log\left[\prod_{i=1}^{n}\left(\frac{1}{(\mu\sigma^2)^{1/\sigma^2}}\cdot\frac{y^{\frac{1}{\sigma^2}-1}e^{-y/\sigma^2\mu}}{\Gamma(1/\sigma^2)}\right)\right]\\
&= \sum_{i=1}^{n}\log\left[\frac{1}{(\mu\sigma^2)^{1/\sigma^2}}\cdot\frac{y^{\frac{1}{\sigma^2}-1}e^{-y/\sigma^2\mu}}{\Gamma(1/\sigma^2)}\right]\\
&= \sum_{i=1}^{n}\left[\log\left(\frac{1}{(\mu\sigma^2)^{1/\sigma^2}}\right)+\log(y^{\frac{1}{\sigma^2}-1}e^{-y/\sigma^2\mu})-\log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right)\right]\\
&= -\frac{n}{\sigma^2}\log(\mu\sigma^2)+\left(\frac{1}{\sigma^2}-1\right)\sum_{i=1}^{n}\log(y_i)-\frac{1}{\sigma^2\mu}\sum_{i=1}^{n}(y_i)\\
&- n\log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right).
\end{aligned} \tag{2.10}$$

Definida a função log-verossimilhança em (2.10), basta encontrar a solução do sistema:

$$\begin{cases}
\dfrac{\partial l(y;\mu,\sigma^2)}{\partial\mu} = 0\\[2ex]
\dfrac{\partial l(y;\mu,\sigma^2)}{\partial\sigma} = 0
\end{cases}$$

- Estimador de μ:

$$\begin{aligned}
\frac{\partial l(y;\mu,\sigma^2)}{\partial\mu} &= \frac{\partial}{\partial\mu}\left(-\frac{n}{\sigma^2}\log(\sigma^2\mu)-\frac{1}{\sigma^2\mu}\sum_{i=1}^{n}(y_i)\right)\\
&= -\frac{n}{\sigma^2\hat{\mu}}+\frac{1}{\sigma^2\hat{\mu}^2}\sum_{i=1}^{n}(y_i)\\
\frac{n}{\sigma^2\hat{\mu}} &= \frac{1}{\sigma^2\hat{\mu}^2}\sum_{i=1}^{n}(y_i)\\
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}(y_i).
\end{aligned}$$

- Estimador de σ^2:

$$
\begin{aligned}
\frac{\partial l(y;\mu,\sigma^2)}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(-\frac{n}{\sigma^2}\log(\mu\sigma^2) + \left(\frac{1}{\sigma^2}-1\right)\sum_{i=1}^{n}\log(y_i) - \frac{1}{\sigma^2\mu}\sum_{i=1}^{n}(y_i) - n\log\left(\Gamma\left(\frac{1}{\sigma^2}\right)\right)\right) \\
&= \frac{2n\hat{\sigma}}{\hat{\sigma}^4}\log(\hat{\sigma}^2\hat{\mu}) - \frac{2n\hat{\sigma}}{\hat{\sigma}^4} + \frac{2n\hat{\sigma}}{\hat{\sigma}^4}\frac{\Gamma'\left(\frac{1}{\hat{\sigma}^2}\right)}{\Gamma'\left(\frac{1}{\hat{\sigma}^2}\right)} + \frac{2\hat{\sigma}}{\hat{\sigma}^4\hat{\mu}}\sum_{i=1}^{n}(y_i) - \frac{2\hat{\sigma}}{\hat{\sigma}^4}\sum_{i=1}^{n}\log(y_i) \\
&= \frac{n}{\hat{\sigma}^4}\log(\hat{\sigma}^2\hat{\mu}) - \frac{n}{\hat{\sigma}^4} + \frac{n}{\hat{\sigma}^4}\frac{\Gamma'\left(\frac{1}{\hat{\sigma}^2}\right)}{\Gamma\left(\frac{1}{\hat{\sigma}^2}\right)} + \frac{1}{\hat{\sigma}^4\hat{\mu}}\sum_{i=1}^{n}(y_i) - \frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}\log(y_i) \\
&= \sum_{i=1}^{n}\left[\frac{1}{\hat{\sigma}^4}\log(\hat{\sigma}^2\hat{\mu}) - \frac{1}{\hat{\sigma}^4} + \frac{1}{\hat{\sigma}^4}\frac{\Gamma'(1/\hat{\sigma}^2)}{\Gamma(1/\hat{\sigma}^2)} + \frac{y_i}{\hat{\sigma}^4\hat{\mu}} - \frac{1}{\hat{\sigma}^4}\log(y_i)\right] \\
&= \sum_{i=1}^{n}\left[\frac{\hat{\mu}\log(\hat{\sigma}^2\hat{\mu}) - \hat{\mu} + \hat{\mu}\Gamma'(1/\hat{\sigma}^2)/\Gamma(1/\hat{\sigma}^2) + y_i - \hat{\mu}\log(y_i)}{\hat{\sigma}^4\hat{\mu}}\right],
\end{aligned}
$$

em que $\dfrac{\Gamma'(1/\sigma^2)}{\Gamma(1/\sigma^2)}$ é denominada função digama, representada por $\psi(1/\sigma^2)$. Por esse motivo, a equação que encontra o estimador do parâmetro σ^2 não pode ser resolvida analiticamente, logo requer o uso de métodos computacionais.

2.3.2 Distribuição Normal

A distribuição normal (NO) é um dos modelos probabilísticos mais importantes da área estatística. Esta é caracterizada por ser uma distribuição apropriada para dados contínuos, simétricos e com suporte em toda reta real. Se uma variável aleatória Y segue distribuição normal com parâmetros μ e σ,ou seja, $Y \sim N(\mu, \sigma^2)$, sua função densidade de probabilidade pode ser definida como em Stasinopoulos, Rigby e Akantziliotou (2008):

$$
f_Y(y|\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \tag{2.11}
$$

em que, $y, \mu \in \Re$ e $\sigma > 0$. A média é dada por $\mathbb{E}(Y) = \mu$ e a variância por $Var(Y) = \sigma^2$, ou seja, a média de Y é μ e o desvio padrão de Y é σ.

A Figura (2.3) na sequência evidência que quanto maior o valor de σ, mais espalhada se torna a distribuição NO, apesar de manter a simetria em torno da média que é uma de suas principais características:

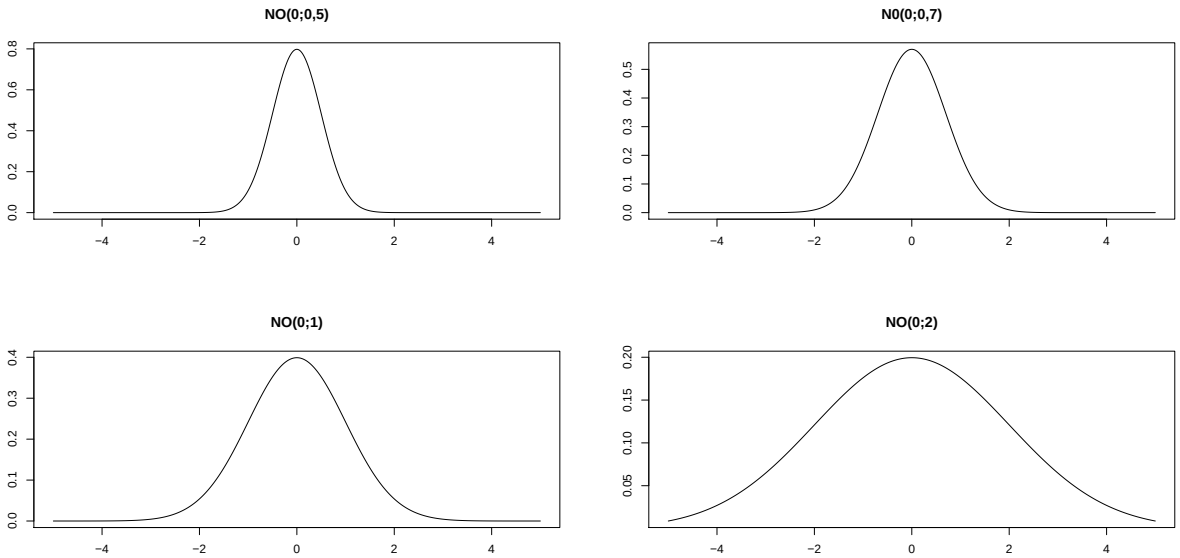

Figura 2.3 – Distribuição Normal para diferentes valores de σ.

2.3.3 Distribuição Log-Normal

Ao considerar uma variável aleatória Y, esta possui distribuição Log-Normal, comumente denotada por $LOGNO(\mu, \sigma^2)$, se a variável X resultante da transformação $\log(Y) = X$ possuir distribuição normal, com média μ e variância e σ^2,respectivamente. Essa transformação permite que as operações com essa variável sempre retorne valores positivos. Sua função densidade de probabilidade é definida em Stasinopoulos, Rigby e Akantziliotou (2008) como:

$$f(y; \mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} exp\left\{ \frac{-(\log(y) - \mu)^2}{2\sigma^2} \right\}, \ \ \forall y, \mu, \lambda > 0. \tag{2.12}$$

Seu respectivo valor esperado e variância são $\mathbb{E}(Y) = w^{\frac{1}{2}} e^{\mu}$ e $Var(Y) = w(w-1)e^{2\mu}$, em que $w = \exp(\sigma^2)$. Concomitante à distribuição GA, é possível observar diferentes formas para a distribuição LOGNO, ficando notório que para valores menores de σ (parâmetro de escala), a distribuição começa mostrar uma certa simetria em torno da média. Tal fato pode ser observado na Figura (2.4):

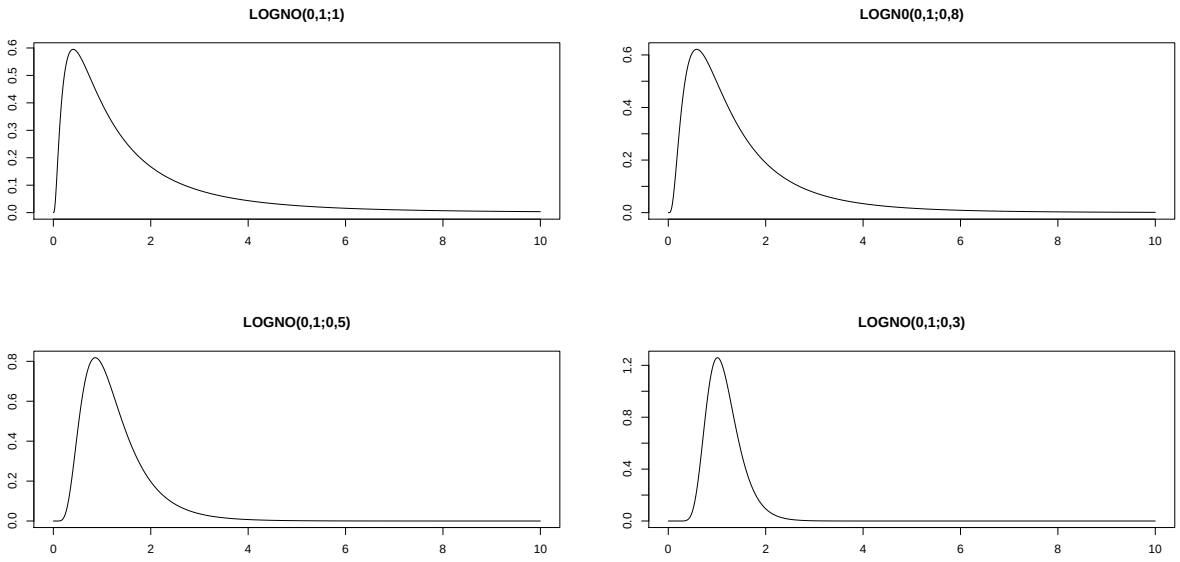

Figura 2.4 – Distribuição Log-normal para diferentes valores de σ.

A distribuição LOGNO é um membro da família exponencial de 2 parâmetros, com parâmetros naturais (η) e estatísticas naturais (T), respectivamente dadas por:

$$\begin{aligned}
f(y;\mu,\sigma^2) &= \exp\left\{\log\left(\frac{1}{y\sqrt{2\pi\sigma^2}}\right) - \frac{(\log(y)-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{\log\left(\frac{1}{y\sqrt{2\pi\sigma^2}}\right) - \frac{\log^2(y) - 2\mu\log(y) + \mu^2}{2\sigma^2}\right\} \\
&= \exp\left\{-\log(y\sqrt{2\pi\sigma^2}) - \frac{\log^2(y) - 2\mu\log(y) + \mu^2}{2\sigma^2}\right\} \\
&= \exp\left\{-\log(y) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\log^2(y)}{2\sigma^2} + \frac{\mu\log(y)}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}.
\end{aligned}$$

em que, $\eta(\mu,\sigma) = \left(\frac{-1}{2\sigma^2}; \frac{\mu}{\sigma^2}\right)$ e $T(Y) = (\log^2(y); \log(y))$.

Assim como para a distribuição GA, é possível definir o respectivo valor esperado e variância da distribuição LOGNO através da família exponencial. Assim, tem-se que:

- Média:

$$\mathbb{E}(Y) = e^{\mu+\frac{\sigma^2}{2}}.$$

- Variância:

$$Var(Y) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1).$$

Para a estimação dos parâmetros, usualmente é utilizado o método da máxima verossimilhança. A função de verossimilhança é definida como:

$$\begin{aligned} L(y_i;\mu,\sigma^2) &= \prod_{i=1}^{n}\left[\frac{1}{y_i\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(\log(y_i)-\mu)^2}{2\sigma^2}\right\}\right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\prod_{i=1}^{n}\left[\frac{1}{y_i}\cdot\exp\left\{-\frac{(\log(y_i)-\mu)^2}{2\sigma^2}\right\}\right]. \end{aligned}$$

Aplicando o logaritmo, encontra-se a função log-verossimilhança:

$$\begin{aligned} l(y_i;\mu,\sigma^2) &= \log\left[\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\prod_{i=1}^{n}\left[\frac{1}{y_i}\cdot exp\left\{-\frac{(\log(y_i)-\mu)^2}{2\sigma^2}\right\}\right]\right] \\ &= -\log(\sqrt{2\pi\sigma^2})^n+\sum_{i=1}^{n}\log\left(\frac{1}{y_i}\right)+\sum_{i=1}^{n}\left[-\frac{(-\log(y_i)-\mu)^2}{2\sigma^2}\right] \\ &= -\frac{n}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{n}\log(y_i)+\sum_{i=1}^{n}\left[-\frac{\log^2(y_i)-2\mu\log(y_i)+\mu^2}{2\sigma^2}\right] \\ &= -\frac{n}{2}\log(2\pi\sigma^2)-\sum_{i=1}^{n}\log(y_i)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\log^2(y_i) \\ &+ \frac{\mu}{\sigma^2}\sum_{i=1}^{n}\log(y_i)-\frac{n\mu^2}{2\sigma^2}. \end{aligned} \tag{2.13}$$

Definida a função log-verossimilhança em (2.13), basta encontrar a solução do sistema a seguir para obter então os estimadores de μ e σ^2, isto é,

$$\begin{cases} \dfrac{\partial l(y;\mu,\sigma^2)}{\partial\mu}=0 \\[2ex] \dfrac{\partial l(y;\mu,\sigma^2)}{\partial\sigma}=0 \end{cases}$$

- Estimador de μ:

$$\begin{aligned} \frac{\partial l(y;\mu,\sigma^2)}{\partial\mu} &= \frac{\partial}{\partial\mu}\left(\frac{\mu}{\sigma^2}\sum_{i=1}^{n}\log(y_i)-\frac{n\mu^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\log(y_i)-\frac{2n\hat{\mu}}{2\sigma^2} \\ \frac{2n\hat{\mu}}{2\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\log(y_i) \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\log(y_i). \end{aligned}$$

- Estimador de σ^2:

$$\begin{aligned}
\frac{\partial l(y;\mu,\sigma^2)}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\log^2(y_i) + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}\log(y_i) - \frac{n\mu^2}{2\sigma^2}\right) \\
&= -\frac{n}{\hat{\sigma}} + \left(\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(\log(y_i)-\hat{\mu})^2\right)' \\
&= -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(\log(y_i)-\hat{\mu})^2 \\
n &= \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(\log(y_i)-\hat{\mu})^2 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\log(y_i)-\hat{\mu})^2.
\end{aligned}$$

A partir disso, é possível fazer as devidas inferências sobre as estimativas dos parâmetros encontrados de acordo com a necessidade e enfoque da pesquisa e/ou análise.

2.3.4 Distribuição Skew Normal

Uma das alternativas que comumente é utilizada nos casos de não normalidade dos dados é a transformação de variáveis. Azzalini e Capitanio (1999), citam que a transformação de variáveis produz resultados razoáveis, porém podem acarretar problemas com a interpretação da variável transformada. Um outro ponto a se ressaltar é que geralmente se transforma cada componente separadamente, tornando a normalidade do todo apenas esperada.

Nesse sentido, a construção de novas famílias de distribuições que possam captar também a assimetria, de forma analiticamente tratável, tornou-se importante. Foi a partir desse cenário que Azzalini (1985) propôs uma nova forma de encontrar distribuições assimétricas, especialmente da família Skew Normal (SN).

É possível encontrar diversas distribuições assimétricas por meio da fórmula proposta por Azzalini (1985):

$$h(y) = 2g(y)\cdot G(w(y)).$$

Aqui, $g(.)$ é uma função densidade simétrica em torno da origem, $G(.)$ é uma função densidade acumulada da função densidade simétrica e $w(.)$ é uma função ímpar qualquer.

A distribuição SN é um exemplo de distribuição oriunda da fórmula de Azzalini. Dada uma variável aleatória Y que segue distribuição SN, ou seja, $Y \sim SN(\lambda)$, esta pode ser definida como:

$$SN(y) = 2\phi(y)\cdot\Phi(\lambda y),\ \ \forall\lambda, y \in \Re,$$

em que, $\phi(.)$ é a função densidade da normal padrão, $\Phi(.)$ é a função densidade acumulada da normal padrão e λ é o parâmetro de assimetria. Contudo, é possível estender esse modelo, introduzindo os parâmetros μ de locação e σ^2 de escala, ou seja, $Y \sim SN(\mu, \sigma^2, \lambda)$. A f.d.p. da variável aleatória Y para esse caso é dada por:

$$f(y; \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \cdot \phi\left(\frac{y-\mu}{\sigma}\right) \cdot \Phi\left(\lambda \frac{y-\mu}{\sigma}\right), \ \ \forall \mu, \lambda \in \Re \text{ e } \sigma > 0. \tag{2.14}$$

Note que essa distribuição possui três parâmetros, μ, σ^2 e λ, em que λ define a assimetria da função (esquerda ou direita) e o quão intensa essa é. Para valores negativos de λ tem-se assimetria negativa e consequentemente para valores positivos, assimetria positiva.

É fácil notar que quando $\lambda = 0$, voltamos ao caso da distribuição normal como já se conhece, ou seja, a f.d.p. da normal com parâmetros μ e σ^2 é um caso particular da distribuição $SN(\mu, \sigma^2, 0)$.

Fixado os valores de μ e σ^2, variando apenas λ é possível perceber como a forma da distribuição muda através dos valores adotados para o parâmetro de assimetria. Quanto mais próximo de zero λ for, mais próximo da distribuição normal será a curva da distribuição SN. Tal fato é observado na Figura (2.5):

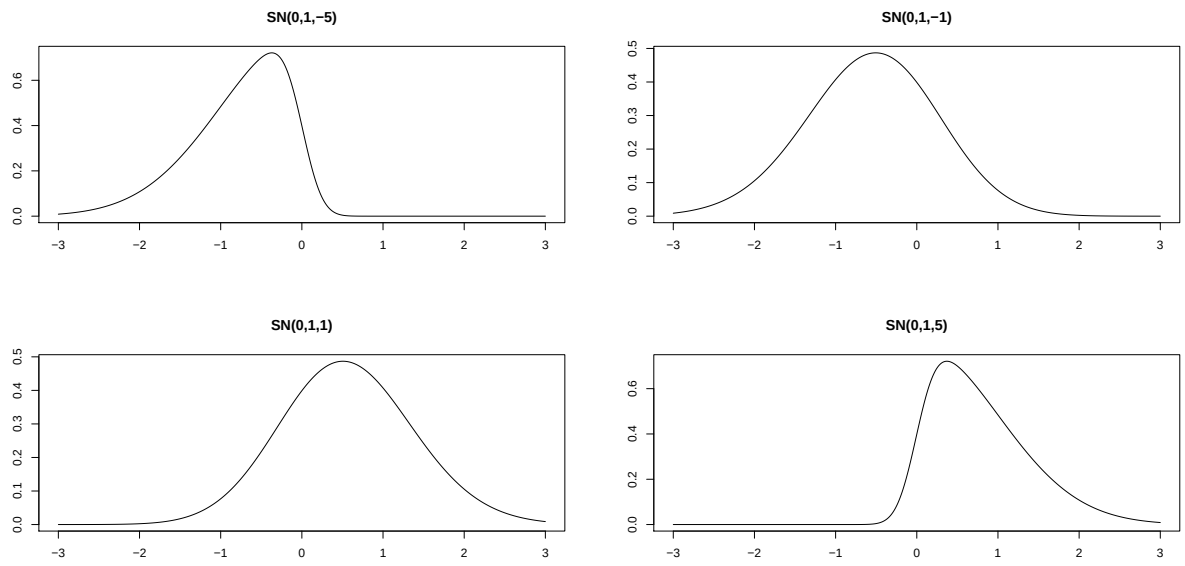

Figura 2.5 – Distribuição Skew Normal para diferentes valores de λ.

De acordo com a equação (2.14), Azzalini (1985) define algumas propriedades da distribuição SN. Para $\rho = \dfrac{\lambda}{\sqrt{1+\lambda^2}}$, tem-se:

- Média:

$$\mathbb{E}(Y) = \mu + \sigma\rho\sqrt{2/\pi}.$$

- Variância:

$$Var(Y) = \sigma^2\left(1 - \frac{2\rho^2}{\pi}\right).$$

- Coeficiente de Assimetria:

$$\gamma_1 = \frac{4-\pi}{2} \cdot \frac{(\sqrt{2/\pi}\rho)^3}{(1-2\rho^2/\pi)^{\frac{3}{2}}}.$$

- Coeficiente de Curtose:

$$\gamma_2 = 2(\pi - 3)\frac{(\sqrt{2/\pi}\rho)^4}{(1-2\rho^2/\pi)^2}.$$

A estimação dos parâmetros pode ser feita por meio do método da máxima verossimilhança ou de momentos. Porém, a equação que encontra os EMV não pode ser encontrada analiticamente, dependendo de métodos computacionais para ser realizada.

A distribuição em questão possui um pacote específico no *Software R*, chamado `sn`. Proposto por Azzalini, sofreu algumas alterações e teve sua última versão atualizada recentemente. O pacote `sn` fornece facilidades para definir e manipular distribuições de probabilidade da família SN e algumas outras relacionadas, aplicar métodos estatísticos para ajuste e diagnóstico de dados, no caso uni e multivariado. Embora o pacote descrito possua várias funções, encontra-se também a implementação da distribuição SN no pacote `gamlss`, optado em ser utilizado, pois incorpora a maioria das distribuições que são utilizadas neste trabalho.

A SN não faz parte da família exponencial de distribuições. Tal fato impediria a modelagem mista por meio dos pacotes convencionais que modelam GLMM. Porém, como já dito, o pacote `gamlss` permite a modelagem mista e possui essa distribuição já implementada, tornando seu uso possível.

Um outro ponto a ser levantado quanto ao uso da SN é o fato de que seu suporte é definido em toda reta real, ou seja, dada uma variável aleatória $Y \sim SN(\mu, \sigma^2, \lambda)$, $y \in \Re$. À vista disso, alguns autores restringem sua utilização apenas para dados pertencentes ao

intervalo $(-\infty, \infty)$. É o caso de Queiroz (2013), que cita em seu trabalho que o uso da SN para dados que não assumem valores negativos pode ser inadequado, sendo necessário procurar outras distribuições que contemple essa característica.

Em contraponto a esse fato, trabalhos como o de Chen, Gupta e Nguyen (2004) fazem uso da modelagem através da SN para dados positivos. Nesse trabalho os autores avaliaram um conjunto de valores referentes à idade de bebês. Dez pares de gêmeos foram acompanhados até que tivessem seu primeiro resfriado a partir da data de seu nascimento. Um outro exemplo é o estudo feito por Guedes et al. (2014), que ajustou modelos de regressão normal com erros assimétricos à dados de alturas de plantas utilizando a distribuição SN.

Dessa forma, não será invalidada as análises realizadas a partir da distribuição SN para o conjunto de dados desse trabalho. Contudo, há na literatura algumas variantes dessa distribuição que poderiam ser alternativas para o caso em que os dados não pertençam à todo intervalo real. Um exemplo, é a distribuição Birnbaum Saunders Skew Normal (BSSN), que têm como suporte o intervalo $\Re^+$. Esta será discutida mais a frente.

2.3.5 Distribuição Skew- t tipo 3

A distribuição Skew-t do tipo 3 (ST3) foi obtida por meio do método proposto por Fernández e Steel (1998) e é uma das cinco versões assimétricas existentes para a distribuição t, definida com quatro parâmetros e com suporte em toda reta real. Vale ressaltar que esta é uma variante da distribuição t, com caldas mais pesadas quando comparada a distribuição SN por exemplo, partindo do pressuposto de que esta também é obtida por meio de um processo específico de assimetrização.

Seja Y uma variável aleatória que segue distribuição ST3, denotada por $ST3(\mu, \sigma^2, \nu, \tau)$. Sua f.d.p. é definida no trabalho de Stasinopoulos, Rigby e Akantziliotou (2008) como:

$$f_Y(y; \mu, \sigma, \nu, \tau) = \frac{c}{\sigma}\left\{1 + \frac{z^2}{\tau}\left[\nu^2 I(y < \mu) + \frac{1}{\nu^2} I(y \geq \mu)\right]\right\}, \tag{2.15}$$

em que $y, \mu \in \Re$, $\sigma, \nu, \tau > 0$, $z = \dfrac{(y-\mu)}{\sigma}$ e $c = 2\nu / \left[\sigma(1+\nu^2) B\left(\frac{1}{2}, \frac{1}{\tau}\right)\tau^{\frac{1}{2}}\right]$.

A Figura (2.6) apresenta o comportamento da distribuição ST3 para diferentes valores de ν, fixado os valores de $\mu = 2, \sigma = 1, \tau = 10$.

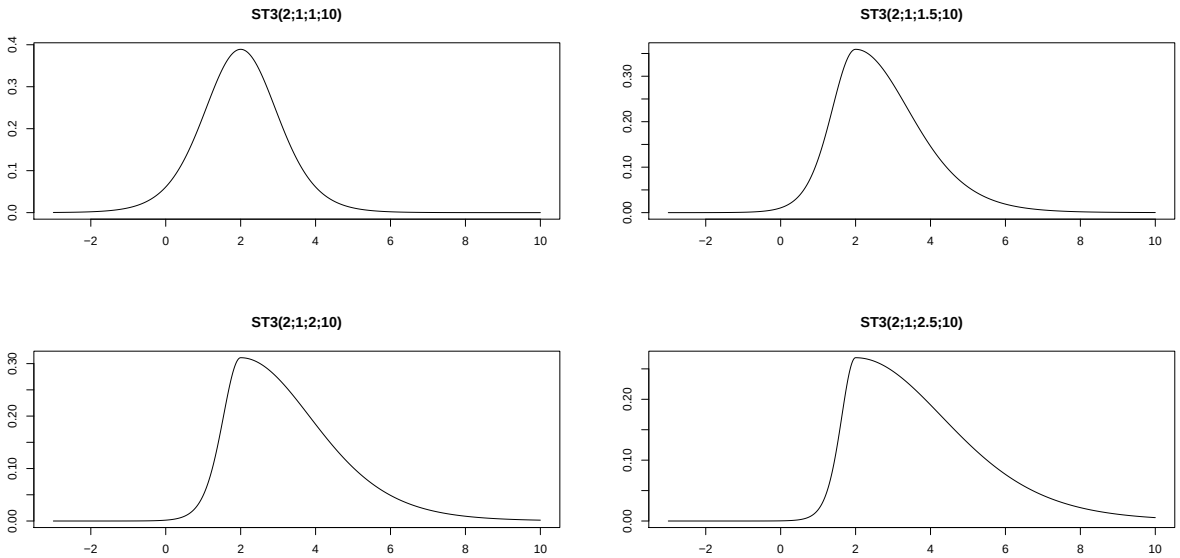

Figura 2.6 – Distribuição Skew-t tipo 3 para diferentes valores de ν.

É notório que com o aumento do valores de ν, os gráficos evidenciam cada vez mais o comportamento assimétrico positivo da curva de densidade. Diferentemente da distribuição SN, tem-se que o parâmetro de assimetria admite somente valores positivos, ou seja, contempla apenas a característica de uma possível assimetria positiva dos dados.

O respectivo valor esperado e variância da distribuição ST3, são dados como em Stasinopoulos, Rigby e Akantziliotou (2008):

- Média:

$$\mathbb{E}(Y) = \mu + \sigma\mathbb{E}(Z)$$

- Variância:

$$Var(Y) = \sigma^2 V(Z),$$

em que,

$$\begin{aligned} \mathbb{E}(Z) &= \frac{2\tau^{\frac{1}{2}}(\nu^2-1)}{(\tau-1)B(\frac{1}{2},\frac{\tau}{2})\nu} \qquad \text{e} \\ \mathbb{E}(Z^2) &= \frac{\tau(\nu^3+\frac{1}{\nu^3})}{(\tau-2)(\nu+\frac{1}{\nu})} \end{aligned}$$

Por ser uma distribuição que tem função de densidade que gera funções que não podem ser resolvidas analiticamente, faz-se uso de recursos computacionais para a estimação de seus parâmetros. Igualmente a distribuição SN, a ST3 não pertence a família

exponencial de distribuições, mas está implementada no pacote `gamlss` do *Software R*, permitindo então que a modelagem mista seja realizada.

É possível fazer uma ligação entre as distribuições SN e ST3, assim como se faz para as distribuições normal e t-Student. Sabe-se que graficamente a distribuição t se assemelha com a normal, sendo simétrica, em forma de sino, porém com caudas mais pesadas (maior variabilidade) quando se tem amostras de tamanhos maiores. O mesmo ocorre quando a comparação é feita entre a SN e ST3. Embora não sejam obtidas através do mesmo método de assimetrização, a ST3 possui caudas mais pesadas do que a SN e quanto maior a amostra, mais essas se assemelham entre si. A partir disso, comparações mais específicas serão realizadas entre essas nos ajustes de cada um dos modelos.

2.3.6 Distribuição Birnbaum Saunders Skew Normal

Proposta por Birnbaum e Saunders (1969), a distribuição Birnbaum Saunders (BS) clássica nasceu a partir de problemas de vibração encontrado em aviões. A família de distribuições BS teve como objetivo inicial modelar o tempo de falha em processos de fadiga.

Seja T uma variável aleatória que segue distribuição BS, ou seja, $T \sim BS(\alpha, \beta)$, sua respectiva f.d.p. é dada por:

$$f_T(t;\alpha,\beta) = \frac{1}{\sqrt{2\pi}} exp\left\{ -\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right\} \frac{t^{\frac{-3}{2}}(t+\beta)}{2\alpha\beta^{\frac{1}{2}}}, \quad \forall t, \alpha, \beta > 0. \tag{2.16}$$

Tomando-se,

$$a_t(\alpha,\beta) = \frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right] \tag{2.17}$$

e

$$\begin{aligned} A_t(\alpha,\beta) &= \frac{\partial}{\partial t} a_t(\alpha,\beta) \\ &= \frac{t^{\frac{-3}{2}}(t+\beta)}{2\alpha\beta^{\frac{1}{2}}}, \end{aligned} \tag{2.18}$$

a f.d.p. pode ser expressa como:

$$f_T(t) = \phi(a_t(\alpha,\beta)) \cdot A_t(\alpha,\beta), \tag{2.19}$$

em que, $\phi(.)$ é a f.d.p da normal padrão, β e α são os parâmetros de escala e forma, respectivamente.

A distribuição BS clássica possui diversas extensões datadas na literatura. Tais extensões se dão principalmente por conta da suposição de normalidade dos dados que esta

exige e muitas das vezes não é satisfeita, compreendendo modelagens que vão desde o ponto de vista bayesiano até o frequentista.

Rieck e Nedelman (1991), formularam e desenvolveram um modelo log-linear para a distribuição BS. O modelo em questão pode ser utilizado para comparar a vida média de várias populações. Além disso, discutiram métodos de análise de dados como o método da máxima verossimilhança e mínimos quadrados. Desmond (1986), investigou a relação entre a distribuição BS e a distribuição normal inversa (IG) e cita que há uma relação íntima entre elas.

No caso em que os dados possuem comportamento assimétrico, uma variante da distribuição BS pode ser a distribuição Birnbaum Saunders Skew Normal (BSSN), que combina a distribuição BS clássica com a distribuição SN. Tal combinação permite modelar dados assimétricos (característica da SN) com suporte nos números reais positivos (característica da BS).

Dessa forma, tem-se que uma variável aleatória T segue distribuição BSSN, ou seja, $T \sim BSSN(\alpha, \beta, \lambda)$ se:

$$f_T(t) = 2\phi(a_t(\alpha, \beta)) \cdot A_t(\alpha, \beta) \cdot \Phi(\lambda a_t(\alpha, \beta)), \tag{2.20}$$

em que, $\phi(.)$ é a f.d.p. da normal padrão, $\Phi(.)$ a acumulada da normal padrão, $\alpha > 0$ é o parâmetro de forma, $\beta > 0$ é o parâmetro de escala e $\lambda \in \Re$ o parâmetro de assimetria. Analogamente, $A_t(\alpha, \beta)$ e $a_t(\alpha, \beta)$ são definidos como em (2.17) e (2.18), respectivamente.

Com relação à uma possível modelagem, encontra-se no *Software R* um pacote denominado `bssn` onde fornece a densidade, função quantil, gerador de números aleatórios, função de confiabilidade, taxa de falhas, função de probabilidade, momentos e algoritmo EM para estimadores de máxima verossimilhança. Além disso, oferece o quantil empírico e envelope gerado para uma determinada amostra, para os três parâmetros de um modelo baseado na distribuição BSSN.

Embora esse pacote exista, nele não há uma implementação que contemple os requisitos para uma modelagem mista. Além disso, diferentemente da distribuição SN discutida anteriormente, a distribuição BSSN não está presente no pacote `gamlss`, logo não permitiu que a modelagem fosse feita como para as demais distribuições.

2.3.7 Distribuição Normal Inversa

Os estudos que envolveram o desenvolvimento da distribuição Normal Inversa (IG) iniciaram-se com os trabalhos de Tweedie (1945). Alguns anos após sua criação, importantes propriedades foram dadas pelo mesmo autor em Tweedie et al. (1957a, 1957b).

A distribuição IG está intimamente ligada a distribuição normal, como sugere seu nome. Uma das grandes diferenças entre elas é dado pelo seu suporte. Enquanto a distribuição normal é definida em todo intervalo real, a distribuição IG é definida apenas para os valores positivos desse mesmo intervalo, ou seja, $\Re^+$.

Assim como as demais distribuições, a IG possui algumas parametrizações e nesse trabalho considera-se o parâmetro $\sigma^2 = \lambda^{-1}$. Dessa forma, dada uma variável aleatória $Y \sim NI(\mu, \lambda)$, essa tem f.d.p. definida como nos trabalhos de Tweedie:

$$f(y; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{ -\frac{\lambda(y-\mu)^2}{2\mu^2 y} \right\}, \ \ \forall x, \mu, \lambda > 0. \tag{2.21}$$

Ao analisar a forma da distribuição IG na Figura (2.7), é fácil ver que fixado o valor de μ e variando os valores de σ, esta aproxima-se da distribuição normal ao passo que σ diminui, ou seja, quando $\lambda \to \infty$.

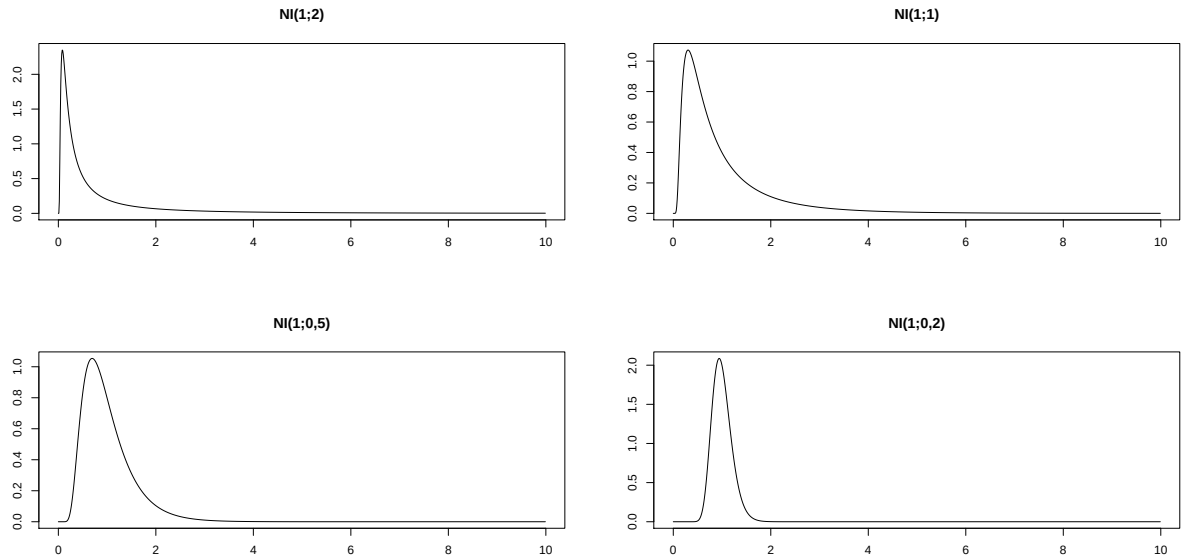

Figura 2.7 – Distribuição Normal Inversa para diferentes valores de σ.

De acordo com Folks e Chhikara (1978), a função densidade vista em (2.21) é unimodal, inclinada positivamente e pertence à família exponencial de distribuições. Portanto, pode ser descrita da seguinte maneira:

$$
\begin{aligned}
f(y;\mu,\lambda) &= \exp\left\{\log\left[\sqrt{\frac{\lambda}{2\pi y^3}}\exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\}\right]\right\} \\
&= \exp\left\{\frac{1}{2}(\log\lambda-\log(2\pi y^3))-\frac{\lambda(y^2-2y\mu+\mu^2)}{2\mu^2 y}\right\} \\
&= \exp\left\{-\frac{\lambda}{2}\left(\frac{y}{\mu^2}-\frac{2}{\mu}+\frac{1}{y}\right)+\frac{1}{2}(\log\lambda-\log(2\pi y^3))\right\} \\
&= \exp\left\{-\frac{\lambda}{2}\left(\frac{y}{\mu^2}-\frac{2}{\mu}\right)+\frac{1}{2}(\log\lambda-\log(2\pi y^3))-\frac{\lambda}{2y}\right\} \\
&= \exp\left\{\lambda\left(-\frac{y}{2\mu^2}+\frac{1}{\mu}\right)-\frac{1}{2}\left[\log\left(\frac{2\pi y^3}{\lambda}\right)+\frac{\lambda}{y}\right]\right\},
\end{aligned}
$$

em que, $\phi=\lambda^{-1}$, $c(y,\phi)=\dfrac{1}{2}\log\left(\dfrac{\lambda}{2\pi y^3}\right)-\dfrac{\lambda}{2y}$, $\theta=-\dfrac{1}{2\mu^2}$ e $b(\theta)=-(-2\theta)^{\frac{1}{2}}$.

O respectivo valor esperado e variância podem ser obtidos por meio da família exponencial e são dados por:

- Média: Considerando $\theta=-\dfrac{1}{2\mu^2}$ e $b(\theta)=-(-2\theta)^{\frac{1}{2}}$, tem-se:

$$
\begin{aligned}
b'(\theta) &= \left[-(-2\theta)^{\frac{1}{2}}\right]' \\
&= \left[-\left(-2\left(-\frac{1}{2\mu^2}\right)\right)^{\frac{1}{2}}\right]' \\
&= \left(\frac{1}{\mu^2}\right)^{-\frac{1}{2}}=\mu.
\end{aligned}
$$

- Variância:

$$
\begin{aligned}
b''(\theta) &= \left[-(-2\theta)^{-\frac{1}{2}}\right]' \\
&= (-2\theta^{-\frac{3}{2}}) \\
&= \left(\frac{1}{\mu^2}\right)^{-\frac{3}{2}}=\mu^3.
\end{aligned}
$$

Assim, $Var(Y)=b''(\theta)\cdot a(\phi)=\mu^3\cdot\dfrac{1}{\lambda}=\dfrac{\mu^3}{\lambda}$.

A estimação de seus parâmetros pode ser realizada pelo método da máxima verossimilhança. Assim, define-se a função de verossimilhança da distribuição IG com parâmetros μ e λ, como:

$$\begin{aligned} L(y_i;\mu,\lambda) &= \prod_{i=1}^{n}\left[\sqrt{\frac{\lambda}{2\pi y_i^3}}\exp\left\{-\frac{\lambda(y_i-\mu)^2}{2\mu^2 y_i}\right\}\right] \\ &= \left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n}\left[y_i^{-\frac{3}{2}}\cdot\exp\left\{-\frac{\lambda(y_i-\mu)^2}{2\mu^2 y_i}\right\}\right]. \end{aligned} \tag{2.22}$$

Aplicando-se o logaritmo, encontra-se a função log-verossimilhança:

$$\begin{aligned} l(y_i;\mu,\lambda) &= \log\left[\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n}\left[y_i^{-\frac{3}{2}}\cdot\exp\left\{-\frac{\lambda(y_i-\mu)^2}{2\mu^2 y_i}\right\}\right]\right] \\ &= \log\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)^n + \sum_{i=1}^{n}\left[\log\left(y_i^{-\frac{3}{2}}\cdot\exp\left\{-\frac{\lambda(y_i-\mu)^2}{2\mu^2 y_i}\right\}\right)\right] \\ &= \log\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)^n + \sum_{i=1}^{n}\log(y_i^{-\frac{3}{2}}) - \sum_{i=1}^{n}\frac{\lambda(y_i-\mu)^2}{2\mu^2 y_i} \\ &= \frac{n}{2}\log(\lambda) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\log(y_i) - \frac{\lambda}{2\mu^2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{y_i} \\ &= \frac{n}{2}\log(\lambda) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\log(y_i) - \frac{n\lambda}{2\mu^2}\sum_{i=1}^{n}\frac{y_i}{n} + \frac{n\lambda}{\mu} - \frac{\lambda}{2}\sum_{i=1}^{n}\frac{1}{y_i} \\ &= \frac{n}{2}\log(\lambda) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\log(y_i) - \frac{n\lambda}{2\mu^2}\bar{y} + \frac{n\lambda}{\mu} - \frac{\lambda}{2}\sum_{i=1}^{n}\frac{1}{y_i}, \end{aligned} \tag{2.23}$$

em que, $\bar{y} = \sum_{i=1}^{n}\frac{y_i}{n}$.

Definida a função log-verossimilhança em (2.23), basta encontrar a solução do sistema a seguir para obter os estimadores de μ e λ:

$$\begin{cases} \dfrac{\partial l(y;\mu,\lambda)}{\partial\mu} = 0 \\[2ex] \dfrac{\partial l(y;\mu,\lambda)}{\partial\lambda} = 0 \end{cases}$$

- Estimador de μ:

$$\begin{aligned}
\frac{\partial l(y;\mu,\lambda)}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-\frac{n\lambda\bar{y}}{2\mu^2}+\frac{n\lambda}{\mu}\right)\\
&= \frac{n\lambda\bar{y}}{\hat{\mu}^3}-\frac{n\lambda}{\hat{\mu}^2}\\
\frac{n\lambda\bar{y}}{\hat{\mu}^3} &= \frac{n\lambda}{\hat{\mu}^2}\\
n\hat{\mu}\lambda &= n\lambda\bar{y}\\
\hat{\mu} &= \bar{y}. \qquad (2.24)
\end{aligned}$$

- Estimador de λ:

$$\begin{aligned}
\frac{\partial l(y;\mu,\lambda)}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left(\frac{n}{2}\log\lambda-\frac{n\lambda}{2\mu^2}-\frac{n\lambda}{\mu}-\frac{\lambda}{2}\sum_{i=1}^{n}\frac{1}{y_i}\right)\\
&= \frac{n}{2\hat{\lambda}}-\frac{n\bar{y}}{2\hat{\mu}^2}+\frac{n}{\hat{\mu}}-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{y_i}\\
\frac{n}{2\hat{\lambda}} &= \frac{n\bar{y}}{2\hat{\mu}^2}-\frac{n}{\hat{\mu}}+\frac{1}{2}\sum_{i=1}^{n}\frac{1}{y_i}\\
\frac{1}{\hat{\lambda}} &= \frac{1}{n}\left[\frac{n\bar{y}}{\hat{\mu}^2}-\frac{2n}{\hat{\mu}}+\sum_{i=1}^{n}\frac{1}{y_i}\right], \text{ substituindo } \hat{\mu}=\bar{y}\\
\frac{1}{\hat{\lambda}} &= \frac{1}{n}\left[-n\frac{1}{\bar{y}}+\sum_{i=1}^{n}\frac{1}{y_i}\right]\\
\frac{1}{\hat{\lambda}} &= \frac{1}{n}\left[\sum_{i=1}^{n}\frac{1}{y_i}-\sum_{i=1}^{n}\frac{1}{\bar{y}}\right]\\
\hat{\lambda} &= \left(\frac{1}{n}\left[\sum_{i=1}^{n}\left(\frac{1}{y_i}-\frac{1}{\bar{y}}\right)\right]\right)^{-1}. \qquad (2.25)
\end{aligned}$$

Encontrados os valores das estimativas dos parâmetros de interesse da distribuição em questão, é possível fazer as devidas inferências de acordo com a perspectiva da pesquisa e/ou análise.

2.3.8 Distribuição t-Student

A distribuição da família TF é adequada para modelar dados leptocúrticos, ou seja, dados com curtose maior do que a distribuição normal. Se Y é uma variável aleatória que segue distribuição t, sua f.d.p. definida em Stasinopoulos, Rigby e Akantziliotou (2008) e

denotada como $Y \sim t(\mu, \sigma, \nu)$, é dada por:

$$f_Y(y|\mu, \sigma, \nu) = \frac{1}{\sigma B\left(\frac{1}{2}; \frac{\nu}{2}\right)\nu^{\frac{1}{2}}} \cdot \left[1 + \frac{(y-\mu)^2}{\sigma^2 \nu}\right]^{-\frac{\nu+1}{2}}, \tag{2.26}$$

em que $y, \mu \in \Re$ e $\sigma, \nu > 0$, onde $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ é a função beta. A média e variância de Y são dadas por $\mathbb{E}(Y) = \mu$ e $Var(Y) = \sigma^2\nu/(\nu - 2)$, com $\nu > 2$.

A Figura (2.8) mostra o comportamento da distribuição TF para diferentes valores de σ. Fica evidente que se assemelha muito a distribuição NO, porém com caudas mais pesadas:

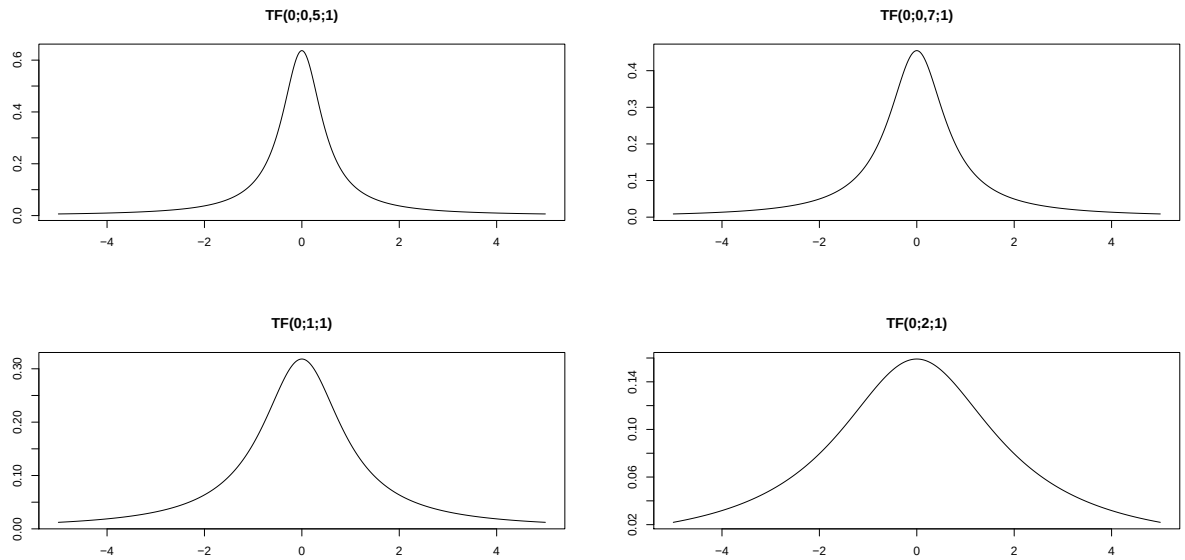

Figura 2.8 – Distribuição t-Student para diferentes valores de σ.

2.3.9 Comparação entre Distribuições

Usualmente adota-se a distribuição GA, IG ou então LOGNO para dados positivos com assimetria à direita. Além de tudo a distribuição GA, por exemplo, está incluída na classe de modelos lineares generalizados, ou seja, pertence a família exponencial de distribuições.

A diferença entre essas distribuições está na transformação de variável. Enquanto na distribuição LOGNO a análise pode ser realizada para a variável transformada $(\log(Y))$, obtendo-se estimativas baseadas em médias geométricas, para a distribuição GA, modela-se o logaritmo de sua média (usando a função de ligação logarítmica) e se obtêm estimativas por meio de médias aritméticas.

Em geral, as análises de dados realizadas a partir dessas distribuições dão origem a resultados similares, levando a mesmas inferências para os modelos (MCCULLAGH;

NELDER, 1989). Assim como a distribuição GA e LOGNO, a distribuição IG também é comumente utilizada em trabalhos que modelam dados positivos assimétricos. Quando comparada a distribuição GA, por exemplo, muitas vezes demonstra melhor qualidade do ajuste quando os dados são ditos demasiadamente assimétricos.

A distribuição SN surgiu da necessidade da construção de famílias de distribuições simétricas que incorporassem também a assimetria. Sua forma mais simples foi proposta por Azzalini (1985), e a partir de então alguns trabalhos sobre o tema foram desenvolvidos a fim de aumentar sua aplicabilidade. Embora não seja tão utilizada como as distribuições GA, IG e LOGNO, se torna uma opção para dados assimétricos positivos no caso em que o parâmetro de assimetria λ assume valores maiores que zero na reta real.

Com relação a distribuição Skew-t tipo 3, esta é uma das cinco variações de assimetrização da distribuição t-Student encontradas no pacote `gamlss` do *Software R*. Possui forte semelhança com a distribuição SN, possuindo caudas mais pesadas. O material encontrado na literatura não é tão vasto quanto para a SN, mas não compromete a sua utilização.

Já a distribuição BSSN, quando comparada com a SN, possui pouquíssima utilização encontrada na literatura. Apesar da distribuição BS ter assumido um importante papel na modelagem do tempo de falha em processos de fadiga, suas variações são geralmente construídas para casos mais específicos e por consequência geram uma menor aplicabilidade.

Quando o assunto são dados com comportamento simétrico, a distribuição normal é uma das grandes referências. Além de ser uma das distribuições mais importantes da área estatística, se caracteriza por ser uma distribuição apropriada para dados contínuos e com suporte nos números reais. Seguindo esse mesmo padrão encontra-se a distribuição t-Student, que se assemelha a normal em algumas de suas características. A respeito da utilização e de trabalhos encontrados na literatura, as duas distribuições em questão apresentam inúmeras possibilidades e com aplicações nas mais diversas áreas.

Capítulo 3

Aplicações

Como mencionado anteriormente, as aplicações a seguir possuem o objetivo em comum de explicar o comportamento da variável resposta diâmetro da lesão, através das possíveis variáveis explicativas folha, DAI e tratamento. Além disso, o foco principal é o de obter resultados quanto aos genótipos mais e menos suscetíveis à doença cancro cítrico, sob a abordagem de modelos mistos utilizando diferentes distribuições de probabilidade. Para a primeira aplicação será utilizado o banco de dados que possui os 16 genótipos de citros, resultado do primeiro experimento, e para a segunda aplicação, os dados referentes aos 6 genótipos derivados do experimento dois.

3.1 Aplicação 1

3.1.1 Análise Descritiva

Para se fazer uma modelagem satisfatória, é imprescindível que uma análise exploratória inicial seja realizada de forma a identificar as principais particularidades dos dados em questão. Como primeiro passo, realizou-se a inspeção visual do histograma da variável resposta diâmetro da lesão.

Esta inspeção se apresenta apenas como um ponto de partida para a análise e não deve servir como critério para escolha da "melhor distribuição"para ser utilizada na modelagem.

Nas Figuras (3.1), (3.2) e (3.3), a linha pontilhada representa o respectivo gráfico de densidade e a linha contínua, as respectivas distribuições de probabilidade, SN, GA, LOGNO, IG e ST3 ajustadas a partir dos parâmetros estimados pelos dados.

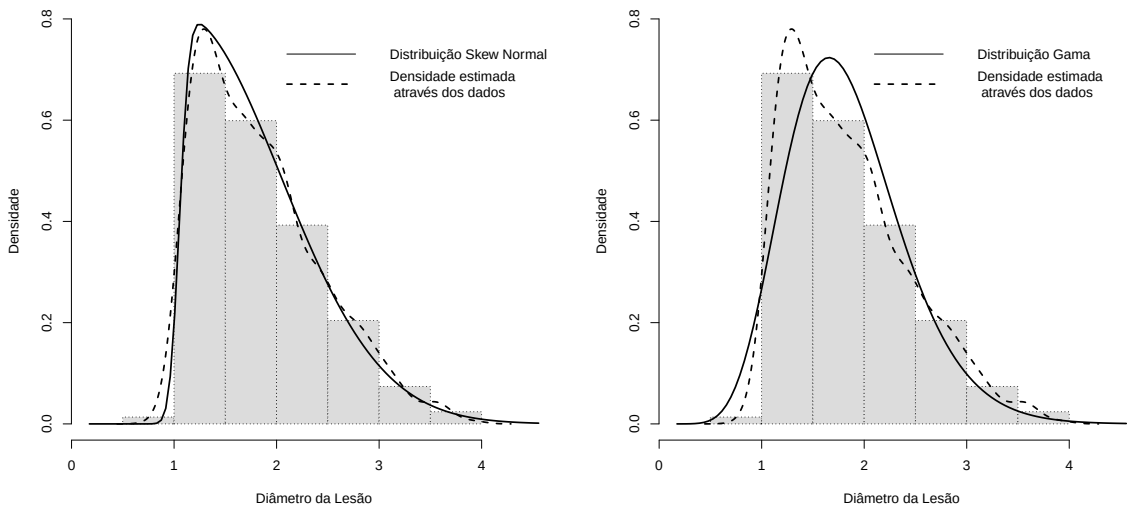

Figura 3.1 – Ajuste das distribuições Skew Normal e Gama à variável resposta diâmetro da lesão.

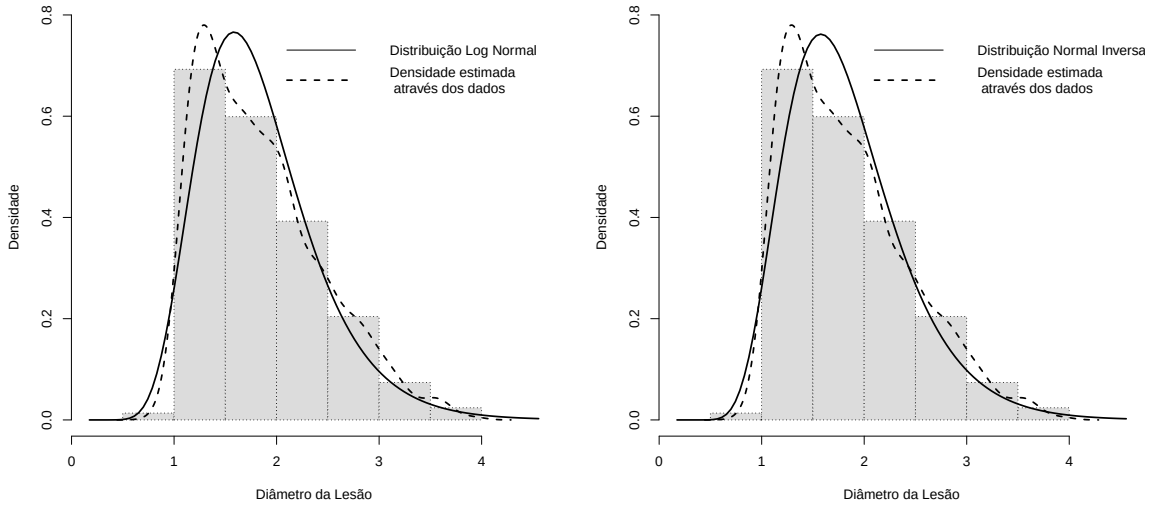

Figura 3.2 – Ajuste das distribuições Log-normal e Normal Inversa à variável resposta diâmetro da lesão.

A Tabela (3.1) apesenta as medidas resumo em relação a cada um dos 16 genótipos de citros avaliados. Ao analisá-la observa-se que os valores médios dos diâmetros das lesões variam consideravelmente entre os genótipos.

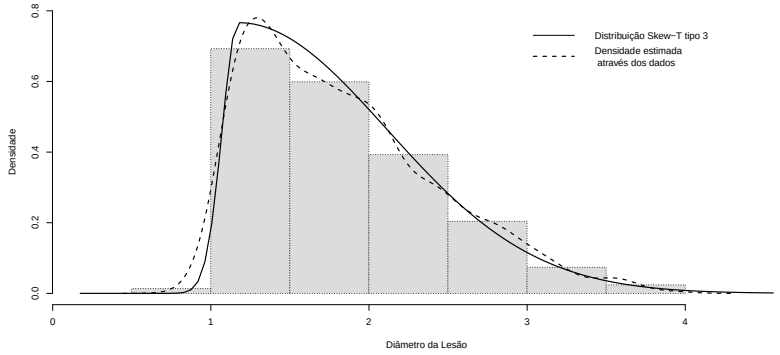

Figura 3.3 – Ajuste da distribuição Skew-t tipo 3 à variável resposta diâmetro da lesão.

Tabela 3.1 – Medidas resumo por genótipo.

Genótipos	Mediana	Min	Máx	Média	DP	CV
Bahia 25-462		1,310	2,990	1,850	0,439	0,237
Morcott 280		**0,960**	**2,860**	**1,605**	**0,465**	**0,290**
Natal 245		1,060	3,930	2,031	0,636	0,313
Natal 261		0,950	2,900	1,750	0,534	0,305
Natal 308		1,060	2,800	1,755	0,450	0,256
Natal M9-324		1,110	2,700	1,715	0,392	0,229
Natal M9-350		1,000	3,450	1,899	0,660	0,347
Pera 329		0,800	2,720	1,683	0,452	0,269
Pera 331		0,940	3,170	1,967	0,663	0,337
Pera 436		**1,030**	**1,420**	**1,234**	**0,114**	**0,092**
Pera 460		1,060	2,800	1,764	0,438	0,248
Rubi 251		**1,010**	**3,690**	**2,376**	**0,846**	**0,356**
Rubi 353		1,240	2,100	1,877	0,227	0,121
Valência 326		1,080	2,450	1,613	0,330	0,204
Westin 16-319		1,100	2,690	1,769	0,433	0,244
Westin 340		**1,070**	**3,910**	**2,502**	**0,778**	**0,311**

O genótipo *Pera 436* apresentou menor diâmetro médio das lesões (1,234), seguido do genótipo *Morcott 280*, com diâmetro médio das lesões de 1,605. A variedade que apresentou maior valor dessa medida foi a *Westin 340*, medindo 2,502. A variabilidade em torno da média segue este mesmo padrão, ou seja, o genótipo com menor desvio padrão é o *Pera 436* e com maior *Westin 340*, com valores de 0,114 e 0,778 respectivamente.

O boxplot apresentado na Figura (3.4) evidencia que para a maioria dos genótipos houve a aparição de valores discrepantes ao menos em um dos dias de avaliação, mostrando que os valores das médias talvez tenham sofrido influência positiva e/ou negativa desses valores, o que torna essa medida sensível e necessitando então de maior cautela ao utilizá-la. É notório também que o diâmetro da lesão cresce com o passar dos DAI em todos os genótipos, embora para alguns esse crescimento aparenta ser menor e para outros maior.

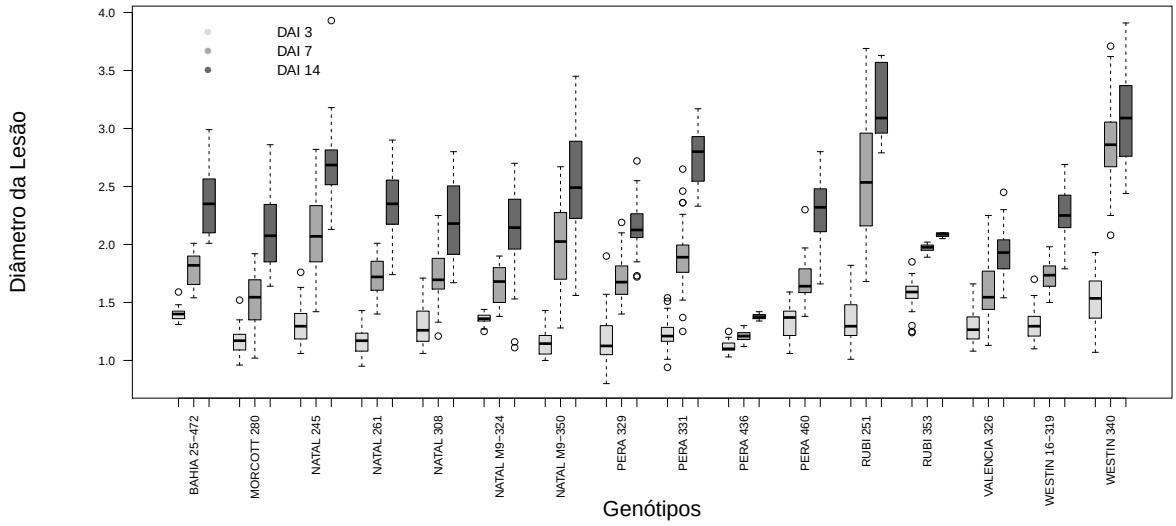

Figura 3.4 – Boxplot da variável diâmetro da lesão para os genótipos em cada DAI.

Além disso, é possível verificar novamente que o genótipo *Pera 436* é o de menor suscetibilidade à doença, quando comparado aos demais, visto que o crescimento do diâmetro das lesões é quase constante através dos DAI, sendo este o que possui também menor variabilidade.

A Figura (3.5) mostra o histograma do diâmetro da lesão em relação a cada um dos DAI.

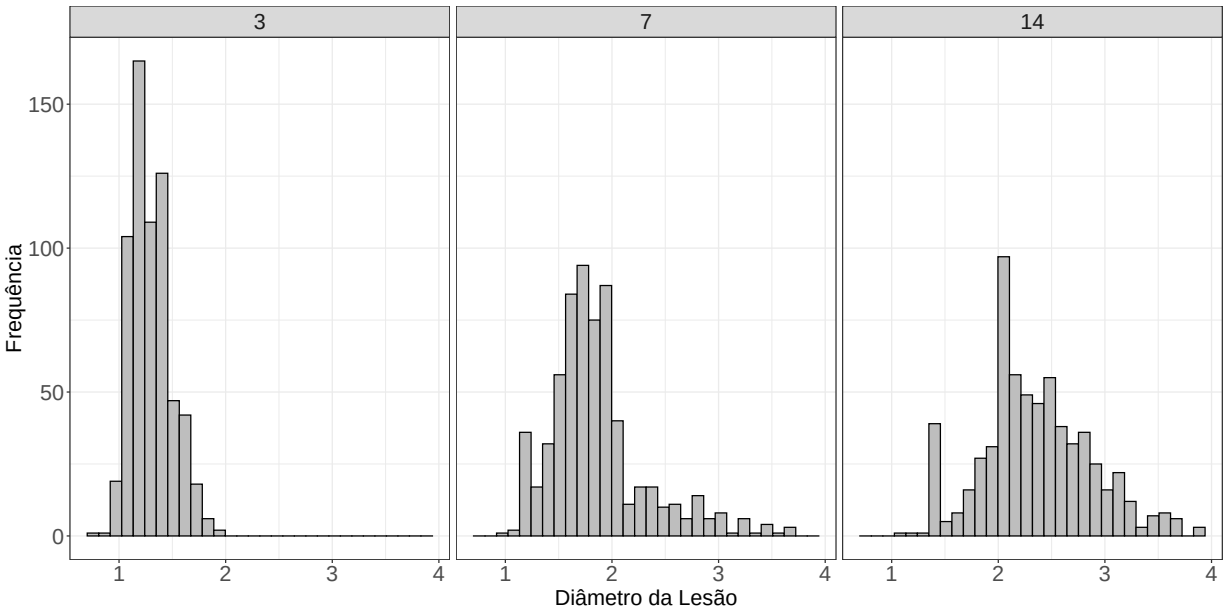

Figura 3.5 – Histograma da variável diâmetro da lesão em cada DAI.

Note que que o histograma corrobora o fato visto no boxplot da Figura (3.4), ao passo que o diâmetro da lesão aumenta em cada um dos DAI. Para o terceiro dia após a inoculação da bactéria, os valores do diâmetro da lesão estão todos concentrados no intervalo de 0 à 2. No sétimo dia esse cenário muda e os valores se dissipam, ficando entre 0 e 4, porém com maior concentração ainda entre os valores de 1 e 2. Já para o último dia de avaliação o diâmetro está entre os valores de 1 à 4, com concentração entre 2 e 3.

De acordo com Singer, Nobre e Rocha (2018), os gráficos de perfis (individuais) são as ferramentas descritivas mais importantes para a análise de dados longitudinais. Eles são essencialmente gráficos de dispersão (com o tempo na abscissa e a resposta na ordenada) em que os pontos associados a uma mesma unidade amostral são unidos por segmentos de reta. Esse tipo de gráfico pode ser usado não só para representar dados longitudinais como também para ajudar na identificação de modelos apropriados para inferência estatística. Nesta análise, o gráfico de perfis relaciona o tamanho da lesão em cada um dos DAI, levando em consideração os genótipos. Como segue:

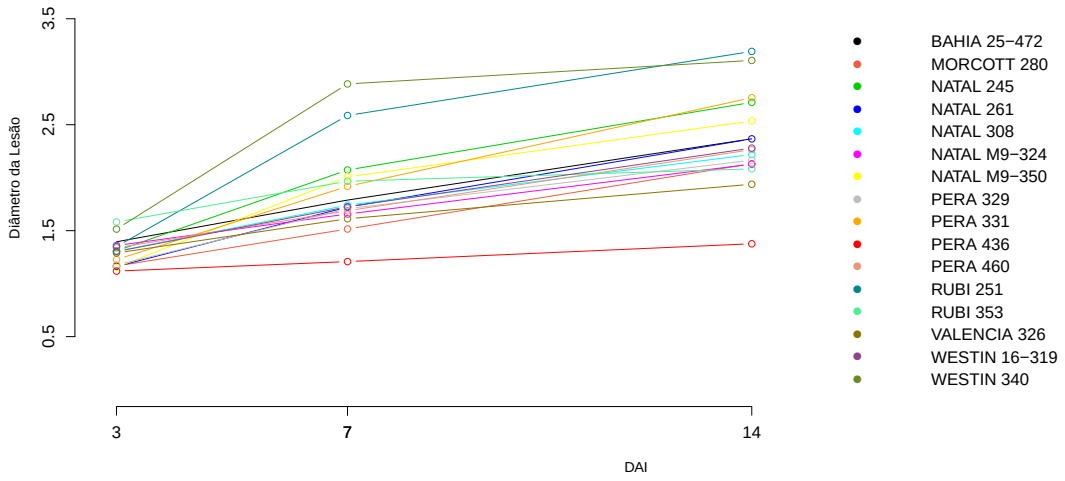

Figura 3.6 – Gráfico de perfis para cada genótipo.

Fica evidente ao examinar a Figura (3.6) que o diâmetro da lesão do genótipo *Pera 436* foi o que apresentou menor crescimento médio com o passar dos dias. Além deste, os genótipos que se destacam com maiores diâmetros são *Rubi 251* e *Westin 340*. Percebe-se que os dois genótipos iniciam com diâmetros médios aparentemente próximos dos demais e logo no segundo dia de avaliação (DAI 7) já mostram um padrão de crescimento elevado das lesões. Até o DAI 7 o genótipo *Westin 340* mostra lesões maiores do que o genótipo *Rubi 251*. Estes papéis se invertem no decorrer dos próximos sete dias, visto que o genótipo *Rubi 251* mostra um decaimento, enquanto o *Westin 340* permanece crescendo até a

última avaliação. Dessa forma, o genótipo que apresenta em média maiores lesões no DAI 14 é o *Rubi 251*.

3.1.2 Ajustes

A variável resposta considerada para os modelos é o diâmetro da lesão. As possíveis variáveis explicativas utilizadas são DAI, tratamento e folha, onde as duas primeiras foram consideradas na parte fixa do modelo e a última na parte aleatória. Os modelos foram ajustados para as cinco distribuições, SN, ST3, IG, GA e LOGNO, considerando a variável resposta y_{ijk}, com $i = 1, 2, ..., 16$, em que i denota os 16 tratamentos, $j = 1, 2, 3$ denota os respectivos dias de avaliação após a inoculação da bactéria (DAI) e $k = 1, 2, ..., 80$ denota o número de folhas destacadas utilizadas no experimento.

Os modelos finais foram ajustados sem alguns outliers, identificados através da análise de resíduos dos modelos iniciais. De fato, observou-se que haviam dois pontos que se mostravam mais discrepantes que os demais, nos resíduos dos cinco modelos propostos. Para a distribuição SN e ST3, além dos pontos mencionados, foi retirado mais um ponto que se mostrava significante do ponto de vista de discrepância. Tais resultados são mostrados nas Figuras (3.10) e (3.11).

Note que não foram apenas estes pontos que se mostraram como outliers nos resíduos dos modelos, porém foram os mais discrepantes. Os demais pontos, apresentaram-se bem próximos dos valores de -3 e 3, não impactando de forma significativa como os mencionados anteriormente.

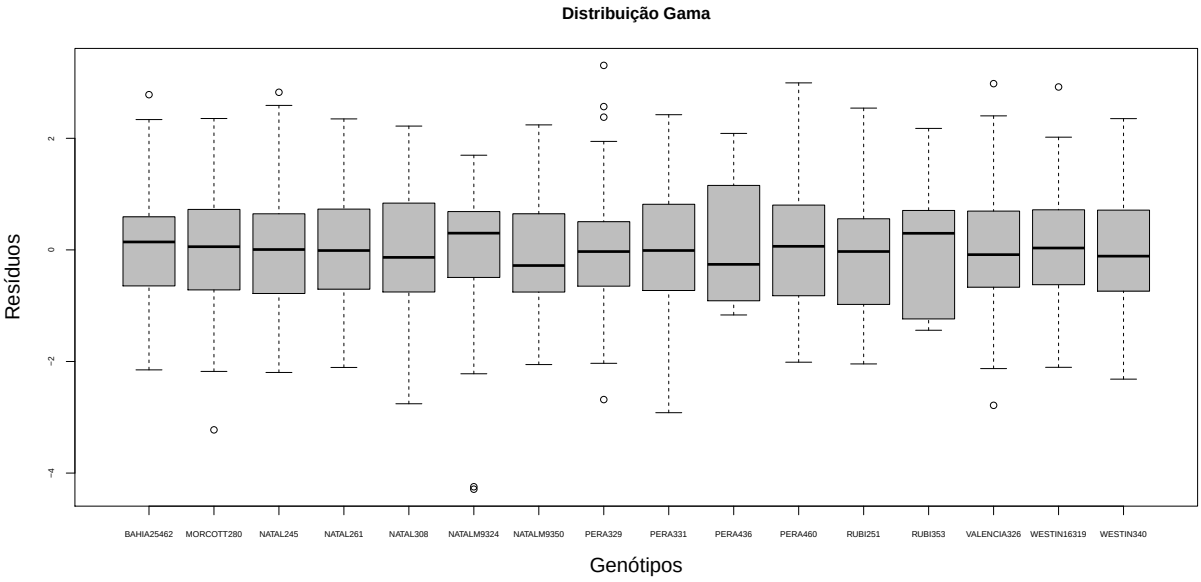

Figura 3.7 – Gráfico de Boxplot dos resíduos do modelo Gama em relação a cada genótipo.

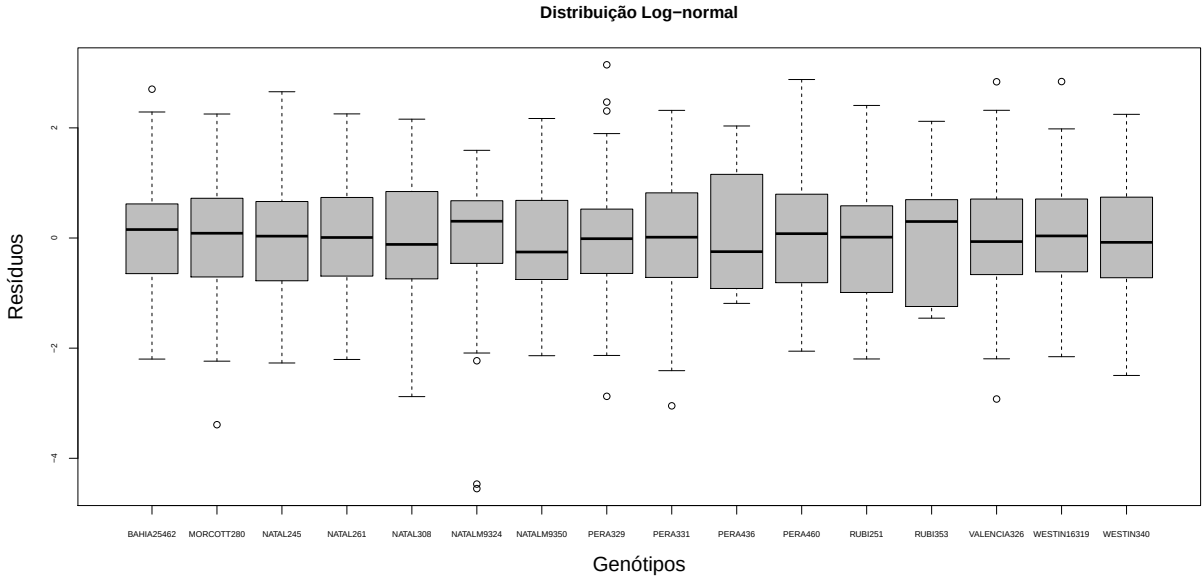

Figura 3.8 – Gráfico de Boxplot dos resíduos do modelo Log-normal em relação a cada genótipo.

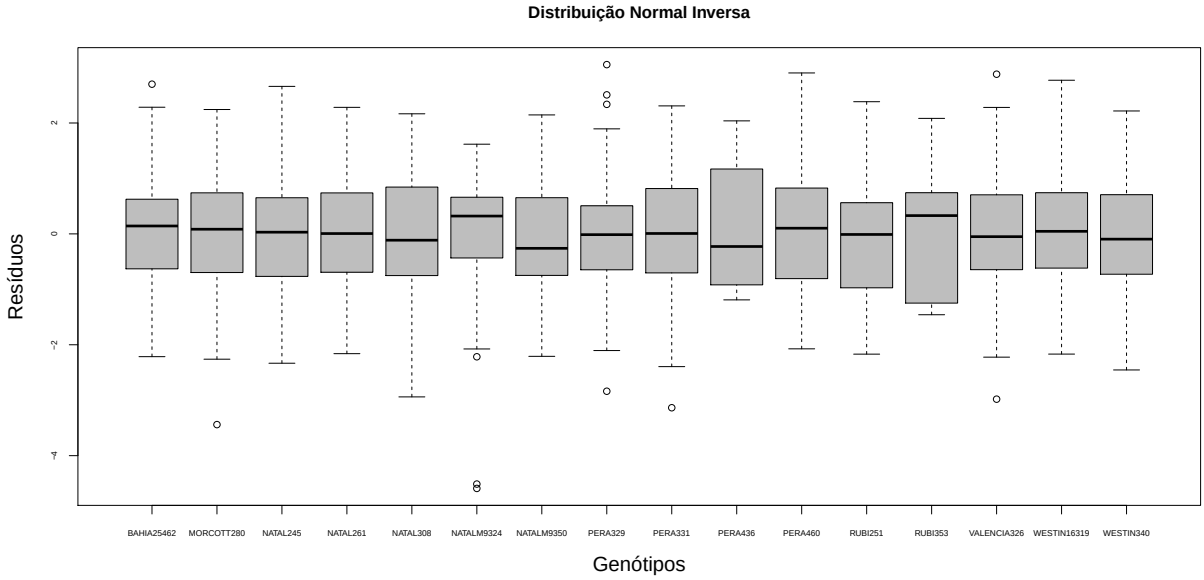

Figura 3.9 – Gráfico de Boxplot dos resíduos do modelo Normal Inverso em relação a cada genótipo.

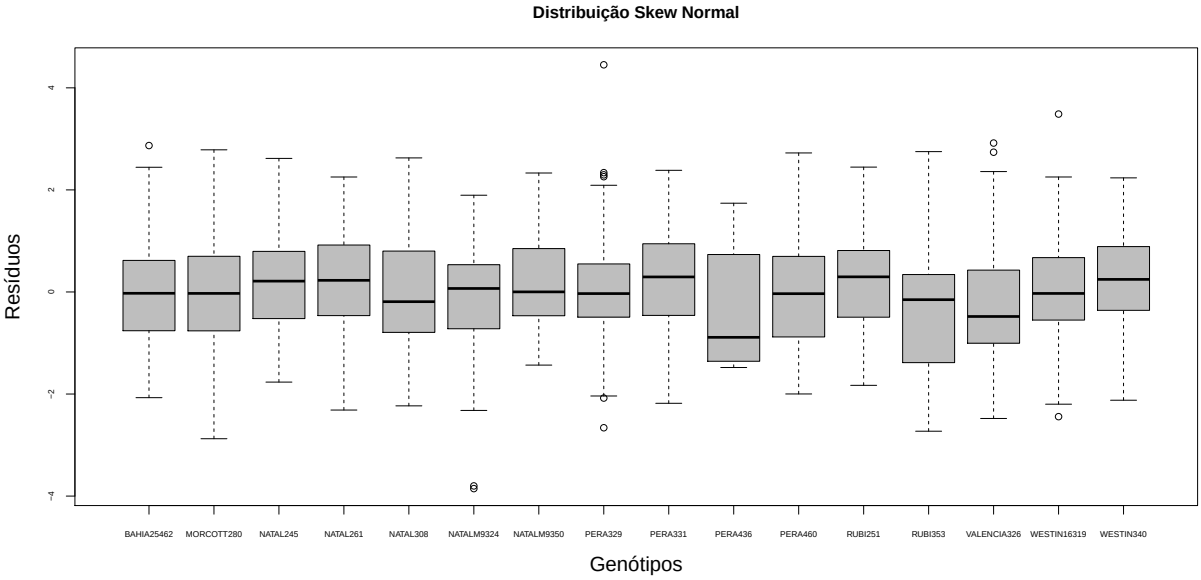

Figura 3.10 – Gráfico de Boxplot dos resíduos do modelo Skew Normal em relação a cada genótipo.

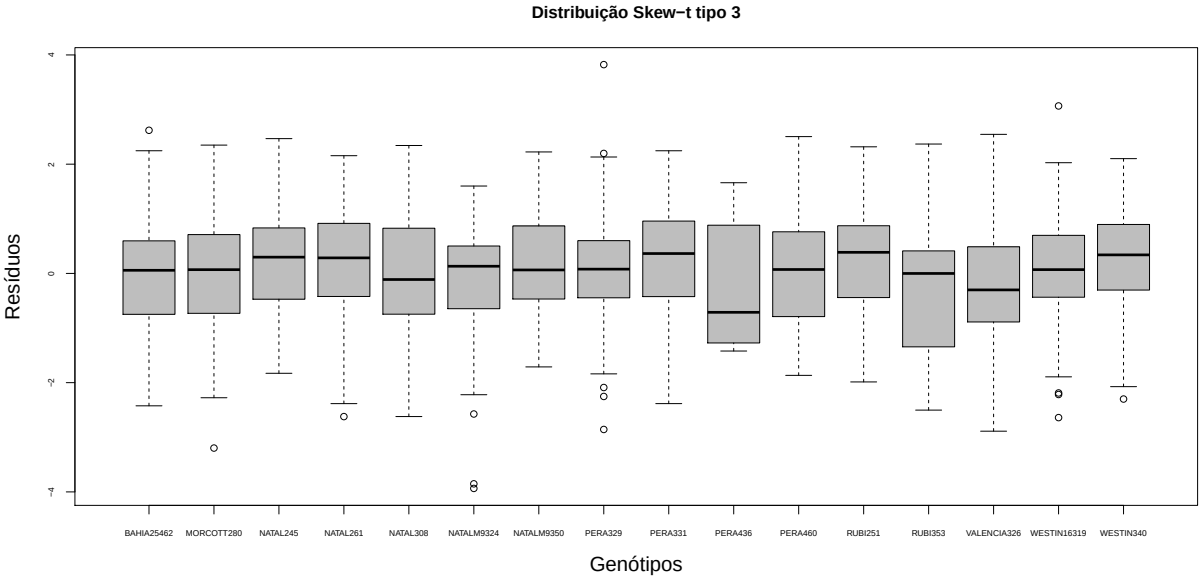

Figura 3.11 – Gráfico de Boxplot dos resíduos do modelo Skew-t tipo 3 em relação a cada genótipo.

Sabe-se que a análise de pontos influentes é de suma importância nas modelagens e que é imprescindível que o pesquisador compreenda o impacto que esses pontos podem causar no resultado final. Como a modelagem em questão é realizada utilizando modelos mistos e esta permite a modelagem de dados desbalanceados, a retirada desses valores seria razoável. Além disso, uma outra questão que foi levada em consideração para essa retirada foi o tamanho da amostra. Obviamente, se o tamanho da amostra fosse considerado pequeno, esses valores poderiam causar problemas ou até inconsistências na modelagem. Neste caso, a amostra é de 1920 observações, que pode ser considerada grande, onde a retirada de dois e/ou três pontos não impactaria de forma severa nos resultados.

Por fim, vale ressaltar que a retirada desses pontos não garante que novos pontos discrepantes apareçam no novo ajuste. Embora tal fato possa ocorrer, a retirada dos anteriores melhorou consideravelmente os gráficos de resíduos e os valores de BIC, AIC e DG para os modelos escolhidos, além de garantir a convergência do modelo IG que anterior às retiradas era falha.

3.1.2.1 Ajuste do Modelo Gama

Seja $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$ uma variável aleatória que segue distribuição GA, o respectivo modelo misto pode ser expresso como:

$$Y_{ij} \sim GA(\mu_{ij}, \sigma^2_{ij}), \text{ com } i = 1, 2, ..., N \text{ e } j = 1, 2, .., n_i,$$

em que, N é o tamanho da amostra e n_i as respectivas observações repetidas. Os efeitos aleatórios, $b_{i\mu}$ e $b_{i\sigma^2}$ são considerados independentes, seguindo distribuição normal com vetor de médias $\mathbf{0}$ e matriz de variâncias e covariâncias $\mathbf{D}_\mu$ e $\mathbf{D}_{\sigma^2}$, respectivamente.

Com isso, escreve-se:

$$Y_{ij}|b_{i\mu}, b_{i\sigma^2} \sim GA(\mu_{ij}, \sigma^2_{ij}),$$

em que, $b_{i\mu} \sim N(0, D_{b_\mu})$ e $b_{i\sigma^2} \sim N(0, D_{b^2_\sigma})$. Além disso, no contexto de GLMM, os parâmetros μ_{ij} e σ^2_{ij}, satisfazem:

$$\begin{aligned} g_\mu &= \eta_{\mu_{ij}} = x^T_{\mu_{ij}}\beta_\mu + Z^T_{\mu_{ij}} b_{i\mu} \\ g_{\sigma^2} &= \eta_{\sigma^2_{ij}} = x^T_{\sigma^2_{ij}}\beta_{\sigma^2} + Z^T_{\sigma^2_{ij}} b_{i\sigma^2}, \end{aligned}$$

sendo os componentes $x^T_{\mu_{ij}}, x^T_{\sigma^2_{ij}}$ e $\beta_\mu, \beta_{\sigma^2}$ referentes à parte fixa do modelo e os componentes $Z^T_{\mu_{ij}}, Z^T_{\sigma^2_{ij}}$ e $b_{i\mu}, b_{i\sigma^2}$ referentes à parte aleatória.

Seja θ_i o efeito fixo do $i-$ésimo tratamento (genótipo), α_j o efeito fixo do $j-$ésimo dia de avaliação após a inoculação da bactéria e γ_k o efeito aleatório da $k-$ésima folha

destacada. Nessas condições, a Tabela (3.3) mostra os seis modelos que foram ajustados e os respectivos valores de AIC, BIC e GD.

Tabela 3.3 – Modelos ajustados para a distribuição Gama.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	3200,76	3211,88	3196,76
2	$\eta_i = \beta_0 + \theta_i trat$	2688,85	2783,38	2654,85
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	187,69	293,33	149,69
4	$\boldsymbol{\eta_{\mathbf{ijk}} = \beta_{\mathbf{0}} + \theta_{\mathbf{i}}\mathbf{trat} + \alpha_{\mathbf{j}}\mathbf{DAI} + \gamma_{\mathbf{0k}}}$	**166,71**	**437,45**	**69,32**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	185,69	285,77	149,69
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	164,71	429,89	69,32

Ao analisar a Tabela (3.3), temos que para o primeiro modelo, ajustou-se apenas o modelo com intercepto, no segundo foi adicionado o efeito fixo dos tratamentos, no terceiro além do efeito dos tratamentos, foi incluído o efeito fixo dos DAI. Somente a partir do modelo quatro que incorporou-se um efeito aleatório, neste caso da variável folha. Testes envolvendo a variável tratamento como efeito fixo e também aleatório foram feitos nos modelos cinco e seis, respectivamente.

Observando os valores dos critérios de seleção, é fácil perceber que é importante para o modelo incorporar o efeito aleatório da folha. Com base nisso, foi testada a inclusão de efeito aleatório também para os tratamentos (modelo cinco), porém sem sucesso, visto que os valores dos critérios de seleção não tiveram grandes mudanças quando comparados ao modelo três, por exemplo. Por fim, o modelo seis que incorpora efeito aleatório tanto para os tratamentos quanto para as folhas, não mostrou mudanças significativas quando comparado ao modelo quatro. Buscando um modelo que melhor explicasse os dados e que fosse o mais parcimonioso possível, a melhor alternativa dentre as expostas foi considerada como sendo o modelo quatro.

O modelo GA final a ser ajustado considerará a função de ligação logarítmica tanto para μ quanto para σ. Assim,

$\log(\mu) = \beta_0$ + θ_1 Morcott 280 + θ_2 Natal 245 + θ_3 Natal 261 + θ_4 Natal 308 + θ_5 Natal M9-324 + θ_6 Natal M9-350 +θ_7 Pera 329 + θ_8 Pera 331 + θ_9 Pera 436 + θ_{10} Pera 460 + θ_{11} Rubi 251 + θ_{12} Rubi 353 + θ_{13} Valência 326 + θ_{14} Westin 16-319 + θ_{15} Westin 340 + θ_{16} Bahia 25-462 + α_1 DAI 3 + α_2 DAI 7 + α_3 DAI 14 + γ_0 FOLHA 1 + γ_0 FOLHA 2 + ... + γ_0 FOLHA 79 + γ_0 FOLHA 80.

$\log(\sigma) = \beta_0$ + θ_1 Morcott 280 + θ_2 Natal 245 + θ_3 Natal 261 + θ_4 Natal 308 + θ_5 Natal

M9-324 + θ_6 Natal M9-350 +θ_7 Pera 329 + θ_8 Pera 331 + θ_9 Pera 436 + θ_{10} Pera 460 + θ_{11} Rubi 251 + θ_{12} Rubi 353 + θ_{13} Valência 326 + θ_{14} Westin 16-319 + θ_{15} Westin 340 + θ_{16} Bahia 25-462 + α_1 DAI 3 +α_2 DAI 7 + α_3 DAI 14 + γ_0 FOLHA 1 + γ_0 FOLHA 2 + ... +γ_0 FOLHA 79 + γ_0 FOLHA 80.

As Tabelas (3.4) e (3.5) na sequência apresentam o resumo do modelo GA para μ e σ considerando como *baseline* a variedade Bahia 25-462 e o DAI 3.

Tabela 3.4 – Estimativas do parâmetro μ para a distribuição Gama.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	0,312	0,009	< 0,01
Morcott 280	$\boldsymbol{\theta_1}$	**-0,151**	**0,014**	**< 0,01**
Natal 245	θ_2	0,089	0,016	< 0,01
Natal 261	θ_3	-0,063	0,014	< 0,01
Natal 308	θ_4	-0,049	0,014	< 0,01
Natal M9-324	θ_5	-0,064	0,011	< 0,01
Natal M9-350	θ_6	0,024	0,019	0,212
Pera 329	θ_7	-0,092	0,014	< 0,01
Pera 331	θ_8	0,046	0,016	< 0,01
Pera 436	$\boldsymbol{\theta_9}$	**-0,384**	**0,015**	**< 0,01**
Pera 460	θ_{10}	-0,049	0,012	< 0,01
Rubi 251	$\boldsymbol{\theta_{11}}$	**0,249**	**0,021**	**< 0,01**
Rubi 353	θ_{12}	0,040	0,013	< 0,01
Valência 326	θ_{13}	-0,125	0,014	< 0,01
Westin 16-319	θ_{14}	-0,045	0,011	< 0,01
Westin 340	$\boldsymbol{\theta_{15}}$	**0,318**	**0,019**	**< 0,01**
DAI 7	α_2	0,291	0,008	< 0,01
DAI 14	α_3	0,539	0,009	< 0,01

A interpretação da Tabela (3.4), evidencia quanto o diâmetro das lesões variam em média em relação ao genótipo Bahia 25-462 e o quanto esse mesmo diâmetro varia em média em relação aos DAI quando comparado do DAI 3. Ao analisar os valores encontrados para as estimativas de cada um dos parâmetros, observa-se que o genótipo *Pera 436* foi o que mostrou menor crescimento médio das lesões no decorrer dos DAI. Os genótipos *Morcott 280* e *Valência 326* seguem a lista de genótipos com crescimento médio das lesões menor que as demais, porém com valores significativamente maiores do que o do genótipo *Pera 436*, por exemplo. Em contrapartida os genótipos que mostraram maiores diâmetros das lesões em média, foram *Westin 340* e *Rubi 251*.

Com relação aos DAI, pode-se dizer que com o passar destes, o diâmetro da lesão tende a aumentar significativamente. Esse crescimento parece se manter constante através dos DAI, não mostrando decaimento em momento algum, ou seja, o maior diâmetro médio

encontrado está presente no DAI 14. Essa análise inicial do parâmetro μ corrobora os resultados apresentados na análise descritiva feita anteriormente.

Tabela 3.5 – Estimativas do parâmetro σ para a distribuição Gama.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	-2,396	0,078	< 0,01
Morcott 280	θ_1	0,511	0,092	< 0,01
Natal 245	θ_2	0,615	0,095	< 0,01
Natal 261	θ_3	0,403	0,094	< 0,01
Natal 308	θ_4	0,471	0,092	< 0,01
Natal M9-324	θ_5	0,143	0,092	0,121
Natal M9-350	θ_6	0,819	0,094	< 0,01
Pera 329	θ_7	0,427	0,095	< 0,01
Pera 331	θ_8	0,601	0,095	< 0,01
Pera 436	θ_9	0,475	0,097	< 0,01
Pera 460	θ_{10}	0,229	0,091	0,012
Rubi 251	θ_{11}	0,925	0,096	< 0,01
Rubi 353	θ_{12}	0,313	0,097	< 0,01
Valência 326	θ_{13}	0,469	0,093	< 0,01
Westin 16-319	θ_{14}	0,666	0,092	0,478
Westin 340	θ_{15}	0,839	0,096	< 0,01
DAI 7	α_2	-0,231	0,050	< 0,01
DAI 14	α_3	-0,063	0,054	0,243

A Tabela (3.5) apresenta os resultados das estimativas do parâmetro σ. Vale ressaltar que esses valores sozinhos não possuem interpretação prática como os valores das estimativas de μ. Porém, sua estimação pode ser uma importante ferramenta para melhorar ainda mais as estimativas de μ.

A partir disso, é importante que se faça a análise dos resíduos do modelo como forma de verificar se este é válido e de fato satisfatório, atendendo as condições para que o ajuste seja considerado adequado.

Na Figura (3.12) são apresentados os gráficos de diagnóstico do modelo GA.

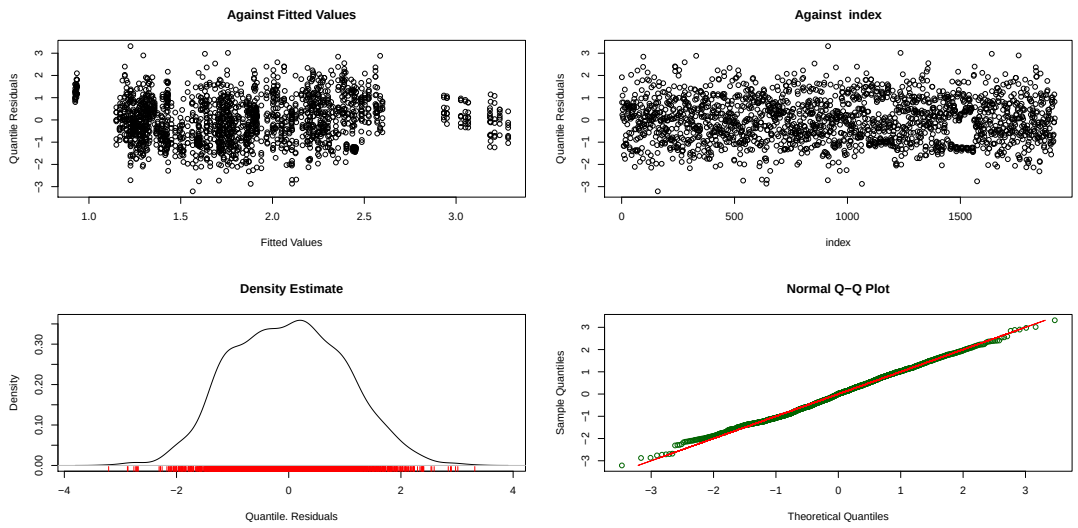

Figura 3.12 – Gráfico de diagnóstico do Modelo Gama.

Além das análises gráficas, observou-se os valores de assimetria, curtose, média e variância do modelo proposto. Os valores obtidos foram satisfatórios visto que a média está próxima de 0, variância próxima de 1, assimetria próxima de 0 e curtose próxima de 3. Esses valores eram esperados desde que os resíduos seguissem distribuição Normal Padrão, o que indica um ajuste adequado. A Tabela 3.6 os evidencia:

Tabela 3.6 – Medidas descritivas dos resíduos do Modelo Gama.

Modelo Gama	
Média	-0,006
Variância	0,998
Coef. Assimetria	0,093
Coef. Curtose	2,649

Por fim, apresenta-se o gráfico Worm-Plot do modelo em questão, onde espera-se que os resíduos, em sua maioria estejam dentro dos limites de confiança e próximos da linha horizontal em torno de zero.

Contudo, levando em consideração todas as ferramentas utilizadas na análise dos resíduos do modelo, pode-se dizer que a distribuição GA é uma opção para o ajuste dos dados, ao passo que obteve-se bons resultados com relação a variabilidade dos resíduos e gráfico q-q plot (informações que podem ser vistas na Figura (3.12)), assim como para os valores das medidas descritivas do modelo, encontradas na Tabela (3.6) e também no gráfico de Worm-Plot da Figura (3.13) onde aparentemente a maioria dos valores estão próximos à linha horizontal em torno de zero e distribuídos dentro dos limites de confiança.

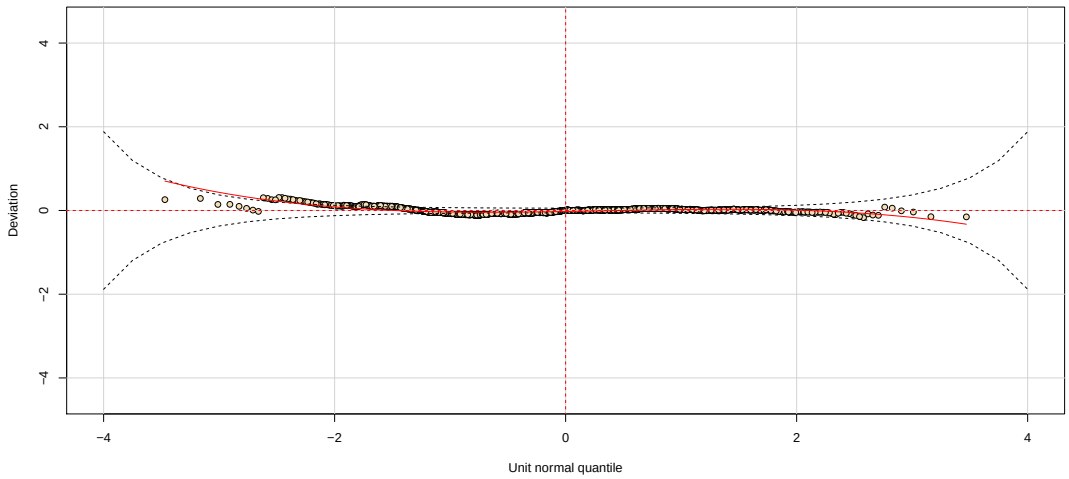

Figura 3.13 – Worm-Plot do Modelo Gama.

3.1.2.2 Ajuste do Modelo Log-Normal

Se $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$ é uma variável aleatória que segue distribuição LOGNO, o respectivo modelo misto pode ser expresso como:

$$Y_i \sim LOGNO(\mu_{ij}, \sigma^2_{ij}), \text{ com } i = 1, 2, ..., N \text{ e } j = 1, 2, .., n_i,$$

em que, N é o tamanho da amostra e n_i as respectivas observações repetidas. Os efeitos aleatórios, $b_{i\mu}$ e $b_{i\sigma^2}$ são considerados independentes, seguindo distribuição normal com vetor de médias $\mathbf{0}$ e matriz de variâncias e covariâncias $\mathbf{D}_\mu$ e $\mathbf{D}_{\sigma^2}$, respectivamente.

Com isso, escreve-se:

$$Y_{ij}|b_{i\mu}, b_{i\sigma^2} \sim LOGNO(\mu_{ij}, \sigma^2_{ij}),$$

em que, $b_{i\mu} \sim N(0, D_{b_\mu})$ e $b_{i\sigma^2} \sim N(0, D_{b^2_\sigma})$.

Além disso, no contexto de GLMM, os parâmetros μ_{ij} e σ^2_{ij}, satisfazem:

$$\begin{aligned} g_\mu &= \eta_{\mu_{ij}} = x^T_{\mu_{ij}}\beta_\mu + Z^T_{\mu_{ij}} b_{i\mu} \\ g_{\sigma^2} &= \eta_{\sigma^2_{ij}} = x^T_{\sigma^2_{ij}}\beta_{\sigma^2} + Z^T_{\sigma^2_{ij}} b_{i\sigma^2}, \end{aligned}$$

onde os componentes $x^T_{\mu_{ij}}, x^T_{\sigma^2_{ij}}$ e $\beta_\mu, \beta_{\sigma^2}$ se referem à parte fixa do modelo e os componentes $Z^T_{\mu_{ij}}, Z^T_{\sigma^2_{ij}}$ e $b_{i\mu}, b_{i\sigma^2}$ se referem à parte aleatória.

Seja θ_i o efeito fixo do $i-$ésimo tratamento (genótipo), α_j o efeito fixo do $j-$ésimo dia de avaliação após a inoculação da bactéria e γ_k o efeito aleatório da $k-$ésima folha

destacada. Nessas condições, a Tabela 3.7 mostra os seis modelos que foram ajustados e os respectivos valores de AIC, BIC e GD.

Tabela 3.7 – Modelos ajustados para a distribuição Log-normal.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	3134,65	3145,77	3130,65
2	$\eta_i = \beta_0 + \theta_i trat$	2716,32	2810,84	2682,32
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	188,19	293,93	150,19
4	$\boldsymbol{\eta_{\mathbf{ijk}} = \beta_{\mathbf{0}} + \theta_{\mathbf{i}}\mathbf{trat} + \alpha_{\mathbf{j}}\mathbf{DAI} + \gamma_{\mathbf{0k}}}$	**165,72**	**442,41**	**66,19**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	186,19	286,27	150,19
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	163,72	434,85	66,19

Ao analisar a Tabela (3.7), é fácil ver que os valores dos critérios de seleção indicam o mesmo comportamento observado no ajuste dos modelos GA, ou seja, a importância do uso de efeito aleatório da variável folha e a variável tratamento sendo testada nos modelos cinco e seis como efeito fixo e também aleatório, porém sem grandes ganhos.

Assim, o modelo LOGNO final a ser ajustado considerará a função de ligação logarítmica tanto para μ quanto para σ seguindo exatamente as mesmas características do modelo GA.

As Tabelas (3.8) e (3.9) na sequência apresentam o resumo do modelo LOGNO para μ e σ considerando como *baseline* a variedade Bahia 25-462 e o DAI 3.

Tabela 3.8 – Estimativas do parâmetro μ para a distribuição Log-normal.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	0,307	0,009	< 0,01
Morcott 280	$\boldsymbol{\theta_1}$	**-0,157**	**0,015**	**< 0,01**
Natal 245	θ_2	0,080	0,016	< 0,01
Natal 261	θ_3	-0,068	0,014	< 0,01
Natal 308	θ_4	-0,055	0,014	< 0,01
Natal M9-324	θ_5	-0,065	0,116	< 0,01
Natal M9-350	θ_6	0,009	0,019	0,646
Pera 329	θ_7	-0,097	0,014	< 0,01
Pera 331	θ_8	0,037	0,016	0,022
Pera 436	$\boldsymbol{\theta_9}$	**-0,389**	**0,015**	**< 0,01**
Pera 460	θ_{10}	-0,052	0,012	< 0,01
Rubi 251	$\boldsymbol{\theta_{11}}$	**0,229**	**0,021**	**< 0,01**
Rubi 353	θ_{12}	0,037	0,013	< 0,01
Valência 326	θ_{13}	-0,130	0,014	< 0,01
Westin 16-319	θ_{14}	-0,046	0,011	< 0,01
Westin 340	$\boldsymbol{\theta_{15}}$	**0,302**	**0,020**	**< 0,01**
DAI 7	α_2	0,295	0,008	< 0,01
DAI 14	α_3	0,541	0,009	< 0,01

Conforme a Tabela (3.8), o genótipo que apresentou menor crescimento médio do diâmetro das lesões no decorrer dos DAI, foi o genótipo *Pera 436*. Os genótipos *Morcott 280* e *Valência 326* seguem a lista de genótipos com crescimento médio das lesões menor que as demais, porém com valores significativamente maiores do que o genótipo *Pera 436*, por exemplo. Os maiores diâmetros das lesões em média, também foram encontrados no genótipo *Westin 340* e *Rubi 251*.

Com relação aos DAI, é notório que com o passar destes o diâmetro médio das lesões tende a aumentar. O crescimento se mostra constante através dos DAI, onde seu maior valor é encontrado no DAI 14.

As estimativas para o parâmetro μ da distribuição LOGNO se apresentam muito próximas as estimativas do mesmo parâmetro para a distribuição GA,os quais se diferem, em sua maioria, somente a partir da segunda casa decimal. Corroborando o fato apresentado anteriormente de que geralmente as estimativas desses modelos se assemelham muito. Assim, os resultados encontrados para o ajuste da distribuição LOGNO reforçam os fatos apresentados na análise descritiva.

Assim como já mencionado, as estimativas de σ apresentadas na Tabela (3.9) sozinhas não possuem interpretações práticas, mas são importantes ao passo que podem melhorar de alguma forma as estimativas de μ.

Tabela 3.9 – Estimativas do parâmetro σ para a distribuição Log-normal.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	-2,400	0,079	< 0,01
Morcott 280	θ_1	0,515	0,092	< 0,01
Natal 245	θ_2	0,614	0,095	< 0,01
Natal 261	θ_3	0,401	0,094	< 0,01
Natal 308	θ_4	0,470	0,092	< 0,01
Natal M9-324	θ_5	0,158	0,092	0,086
Natal M9-350	θ_6	0,814	0,094	< 0,01
Pera 329	θ_7	0,423	0,096	< 0,01
Pera 331	θ_8	0,608	0,095	< 0,01
Pera 436	θ_9	0,472	0,097	< 0,01
Pera 460	θ_{10}	0,231	0,092	0,012
Rubi 251	θ_{11}	0,932	0,097	< 0,01
Rubi 353	θ_{12}	0,319	0,097	< 0,01
Valência 326	θ_{13}	0,468	0,093	< 0,01
Westin 16-319	θ_{14}	0,064	0,093	0,487
Westin 340	θ_{15}	0,848	0,097	< 0,01
DAI 7	α_2	-0,219	0,051	< 0,01
DAI 14	α_3	-0,058	0,055	0,297

A análise gráfica dos resíduos do modelo LOGNO é apresentada na Figura (3.14):

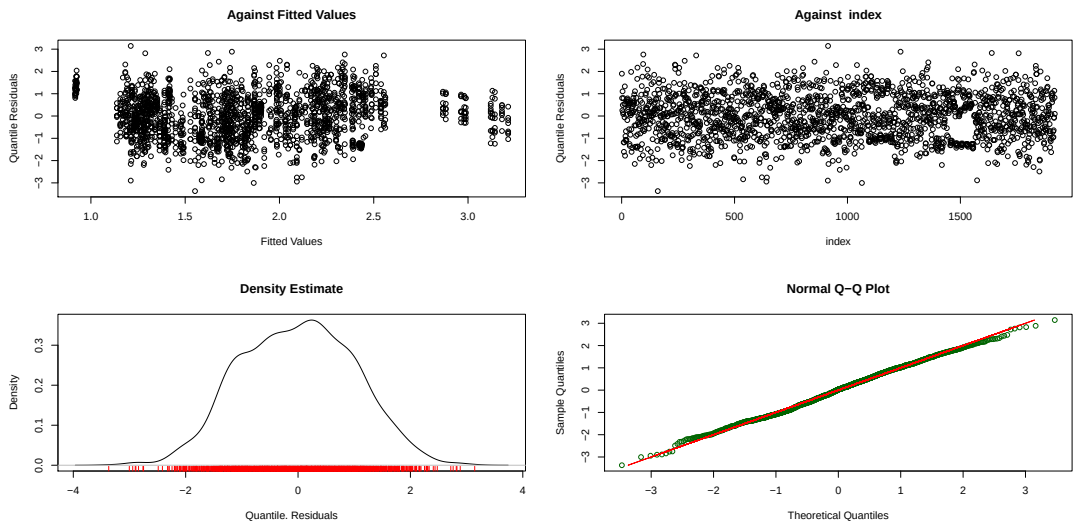

Figura 3.14 – Gráfico de diagnóstico do Modelo Log-normal.

O gráfico Worm-Plot do modelo é dado na Figura (3.15), onde se espera que os resíduos estejam, em sua maioria, dentro dos limites de confiança e próximos da linha horizontal em torno de zero.

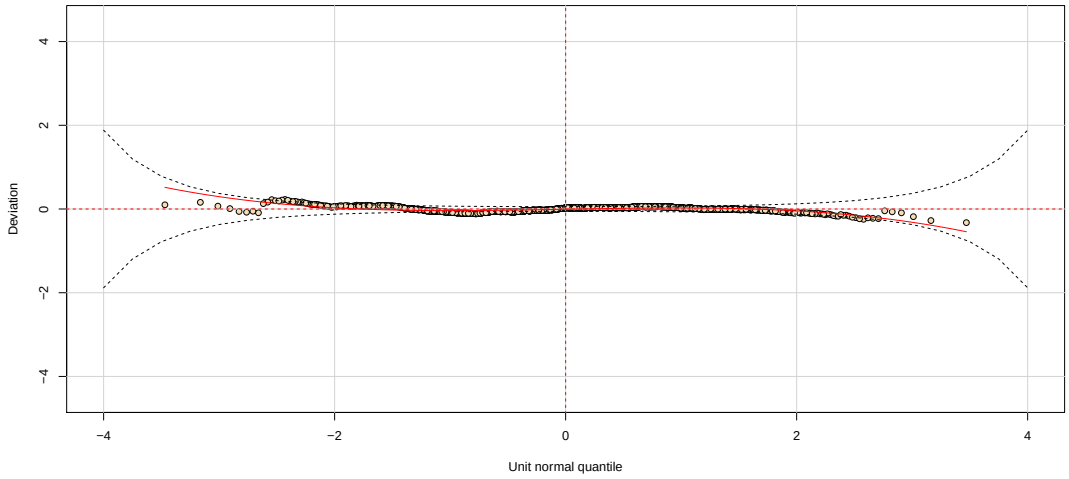

Figura 3.15 – Worm-Plot do Modelo Log-normal.

Além da análise gráfica, os valores de assimetria e curtose são próximos de 0 e 3, respectivamente, bem como os valores da média são próximos de 0 e variância próximos de 1. Tal fato evidencia que os resíduos seguem distribuição Normal Padrão, indicando um ajuste adequado. Os valores reais são dados na Tabela (3.10).

Tabela 3.10 – Medidas descritivas dos resíduos do Modelo Log-normal.

Modelo Log-normal	
Média	-0,006
Variância	1,000
Coef. Assimetria	-0,006
Coef. Curtose	2,629

Assim, levando em consideração todos os recursos utilizados na análise de resíduos do modelo mencionados anteriormente, Figuras (3.14) e (3.15) e Tabela (3.10), pode-se dizer que a distribuição LOGNO se mostra como uma opção para o ajuste dos dados, ao passo que apresenta resultados satisfatórios.

3.1.2.3 Ajuste do Modelo Normal Inverso

Seja $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$ uma variável aleatória que segue distribuição IG, seu respectivo modelo misto pode ser expresso na forma:

$$Y_i \sim IG(\mu_{ij}, \sigma^2_{ij}), \text{ com } i = 1, 2, ..., N \text{ e } j = 1, 2, .., n_i,$$

em que, N é o tamanho da amostra e $n_i$as respectivas observações repetidas. Os efeitos aleatórios, $b_{i\mu}$ e $b_{i\sigma^2}$ são considerados independentes, seguindo distribuição normal com vetor de médias 0 e matriz de variâncias e covariâncias D_μ e D_{σ^2}, respectivamente.

Com base nisso, tem-se:

$$Y_{ij}|b_{i\mu}, b_{i\sigma^2} \sim IG(\mu_{ij}, \sigma^2_{ij}),$$

em que, $b_{i\mu} \sim N(0, D_{b_\mu})$ e $b_{i\sigma^2} \sim N(0, D_{b^2_\sigma})$.

No contexto de GLMM, os parâmetros μ_{ij} e σ^2_{ij}, satisfazem:

$$\begin{aligned} g_\mu &= \eta_{\mu_{ij}} = x^T_{\mu_{ij}}\beta_\mu + Z^T_{\mu_{ij}} b_{i\mu} \\ g_{\sigma^2} &= \eta_{\sigma^2_{ij}} = x^T_{\sigma^2_{ij}}\beta_{\sigma^2} + Z^T_{\sigma^2_{ij}} b_{i\sigma^2}, \end{aligned}$$

sendo os componentes $x^T_{\mu_{ij}}, x^T_{\sigma^2_{ij}}$ e $\beta_\mu, \beta_{\sigma^2}$ referentes à parte fixa do modelo e os componentes $Z^T_{\mu_{ij}}, Z^T_{\sigma^2_{ij}}$ e $b_{i\mu}, b_{i\sigma^2}$ referentes à parte aleatória.

Seja θ_i o efeito fixo do $i-$ésimo tratamento (genótipo), α_j o efeito fixo do $j-$ésimo dia de avaliação após a inoculação da bactéria e γ_k o efeito aleatório da $k-$ésima folha destacada. Nessas condições, a Tabela (3.11) mostra os seis modelos que foram ajustados e os respectivos valores de AIC, BIC e GD.

Tabela 3.11 – Modelos ajustados para a distribuição Normal Inversa.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	3120,91	3132,03	3116,91
2	$\eta_i = \beta_0 + \theta_i trat$	2620,76	2715,28	2586,76
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	218,68	324,32	180,68
4	$\boldsymbol{\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k}}$	**204,12**	**454,64**	**114,01**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	216,68	316,76	180,68
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	202,12	447,08	114,01

Observe que no primeiro modelo, ajustou-se apenas o modelo com intercepto, no segundo foi adicionado o efeito fixo dos tratamentos, no terceiro além do efeito dos tratamentos, foi incluído o efeito fixo dos DAI. A partir do modelo quatro, incorporou-se um efeito aleatório, neste caso da variável folha. Testes envolvendo a variável tratamento como efeito fixo e também aleatório foram feitos nos modelos cinco e seis, respectivamente.

Através dos valores dos critérios de seleção, é fácil perceber a importância do efeito aleatório da folha ser incorporado no modelo. Com base nisso, foi testada a inclusão de

efeito aleatório também para os tratamentos no modelo cinco, porém sem sucesso, visto que os valores dos critérios de seleção não tiveram grandes mudanças quando comparados ao modelo três, por exemplo. Finalmente, o último modelo incorpora efeito aleatório tanto para os tratamentos quanto para as folhas, não apresentando mudanças significativas quando comparado ao modelo quatro. Buscando um modelo que melhor explicasse os dados e que fosse o mais parcimonioso possível, a melhor alternativa dentre as expostas foi considerada como sendo o modelo quatro.

O modelo IG final a ser ajustado considerará a função de ligação logarítmica tanto para μ quanto para σ e possuirá as mesmas características dos anteriores.

As Tabelas (3.12) e (3.13) na sequência apresentam o resumo do modelo para μ e σ considerando como *baseline* a variedade Bahia 25-462 e o DAI 3.

Tabela 3.12 – Estimativas do parâmetro μ para a distribuição Normal Inversa.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	0,313	0,009	< 0,01
Morcott 280	$\boldsymbol{\theta_1}$	**-0,150**	**0,015**	**< 0,01**
Natal 245	θ_2	0,092	0,016	< 0,01
Natal 261	θ_3	-0,061	0,014	< 0,01
Natal 308	θ_4	-0,049	0,014	< 0,01
Natal M9-324	θ_5	-0,065	0,011	< 0,01
Natal M9-350	θ_6	0,027	0,019	0,156
Pera 329	θ_7	-0,091	0,014	< 0,01
Pera 331	θ_8	0,049	0,016	< 0,01
Pera 436	$\boldsymbol{\theta_9}$	**-0,387**	**0,015**	**< 0,01**
Pera 460	θ_{10}	-0,050	0,012	< 0,01
Rubi 251	$\boldsymbol{\theta_{11}}$	**0,253**	**0,021**	**< 0,01**
Rubi 353	θ_{12}	0,039	0,013	< 0,01
Valência 326	θ_{13}	-0,126	0,014	< 0,01
Westin 16-319	θ_{14}	-0,045	0,011	< 0,01
Westin 340	$\boldsymbol{\theta_{15}}$	**0,323**	**0,020**	**< 0,01**
DAI 7	α_2	0,289	0,008	< 0,01
DAI 14	α_3	0,537	0,009	< 0,01

A interpretação da Tabela (3.12), evidencia quanto o diâmetro das lesões variam em média em relação ao genótipo Bahia 25-462 e o quanto esse mesmo diâmetro varia em média em relação aos DAI quando comparado do DAI 3.

Com relação aos valores encontrados para as estimativas de cada um dos parâmetros, nota-se que o genótipo *Pera 436* foi o que mostrou menor crescimento médio das lesões no decorrer dos DAI. Os genótipos *Morcott 280* e *Valência 326* seguem a lista de genótipos com crescimento médio das lesões menor quando comparado aos demais. Por outro lado, os maiores diâmetros das lesões em média, foram encontrados nos genótipos *Westin 340* e *Rubi 251*.Referente aos DAI, o diâmetro da lesão tende a aumentar significativamente no decorrer das avaliações. Esse crescimento parece se manter constante, não mostrando decaimento. Assim o maior diâmetro médio é encontrado no DAI 14.

Da mesma maneira que para as distribuições GA e LOGNO, as estimativas encontradas para o parâmetro μ são muito próximas umas das outras, reforçando mais uma vez as considerações feitas na análise descritiva inicial.

A Tabela (3.13) indica os valores das estimativas do parâmetro σ onde este não possui interpretação quando analisado de forma isolada, assim como mencionado para as demais distribuições. Porém, se bem estimado, auxilia na melhora da estimação do parâmetro μ, o que é de grande interesse.

Tabela 3.13 – Estimativas do parâmetro σ para a distribuição Normal Inversa.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	-2,52649	0,07825	< 0,01
Morcott 280	θ_1	0,59771	0,09208	< 0,01
Natal 245	θ_2	0,57731	0,09507	< 0,01
Natal 261	θ_3	0,43897	0,09408	< 0,01
Natal 308	θ_4	0,49890	0,09212	< 0,01
Natal M9-324	θ_5	0,19048	0,09210	0,039
Natal M9-350	θ_6	0,81232	0,09414	< 0,01
Pera 329	θ_7	0,47215	0,09513	< 0,01
Pera 331	θ_8	0,59201	0,09490	< 0,01
Pera 436	θ_9	0,63956	0,09773	< 0,01
Pera 460	θ_{10}	0,25450	0,09167	< 0,01
Rubi 251	θ_{11}	0,81439	0,09655	< 0,01
Rubi 353	θ_{12}	0,28770	0,09708	< 0,01
Valência 326	θ_{13}	0,53121	0,09286	< 0,01
Westin 16-319	θ_{14}	0,08569	0,09261	0,355
Westin 340	θ_{15}	0,69729	0,09640	< 0,01
DAI 7	α_2	-0,41265	0,05008	< 0,01
DAI 14	α_3	-0,36216	0,05406	< 0,01

Contudo, é de suma importância que se faça a análise dos resíduos do modelo como

forma de verificar se este é válido e de fato satisfatório, atendendo as condições para que o ajuste seja considerado adequado.

Na Figura (3.16) são apresentados os gráficos de diagnóstico do modelo IG.

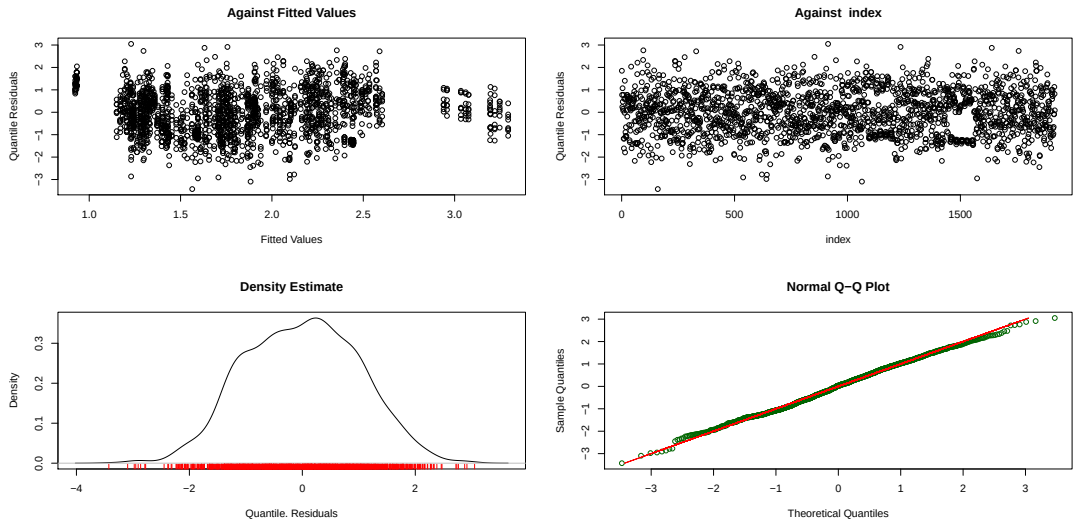

Figura 3.16 – Gráfico de diagnóstico do Modelo Normal Inverso.

Além das análises gráficas, observou-se os valores de assimetria, curtose, média e variância do modelo proposto. Os valores obtidos foram satisfatórios visto que a média está próxima de 0, variância próxima de 1, assimetria próxima de 0 e curtose próxima de 3. Esses valores eram esperados desde que os resíduos seguissem distribuição Normal Padrão, indicando um ajuste adequado. A Tabela (3.14) os evidencia:

Tabela 3.14 – Medidas descritivas dos resíduos do Modelo Normal Inverso.

Modelo Normal Inverso	
Média	-0,007
Variância	0,997
Coef. Assimetria	-0,011
Coef. Curtose	2,649

Por último, apresenta-se na Figura (3.17) o gráfio Worm-Plot e espera-se que os resíduos, em sua maioria estejam dentro dos limites de confiança e próximos da linha horizontal em torno de zero.

Desta forma, pode-se dizer ao explorar os gráficos das Figuras (3.16) e (3.17), além da Tabela (3.14) que os resíduos do modelo são satisfatórios, vistas as informações obtidas para a variabilidade, gráfico q-q plot, gráfico Worm-Plot e medidas descritivas que indicam que estes seguem distribuição próxima da normal padrão, tornando assim a distribuição IG também uma opção para a modelagem dos dados dessa aplicação.

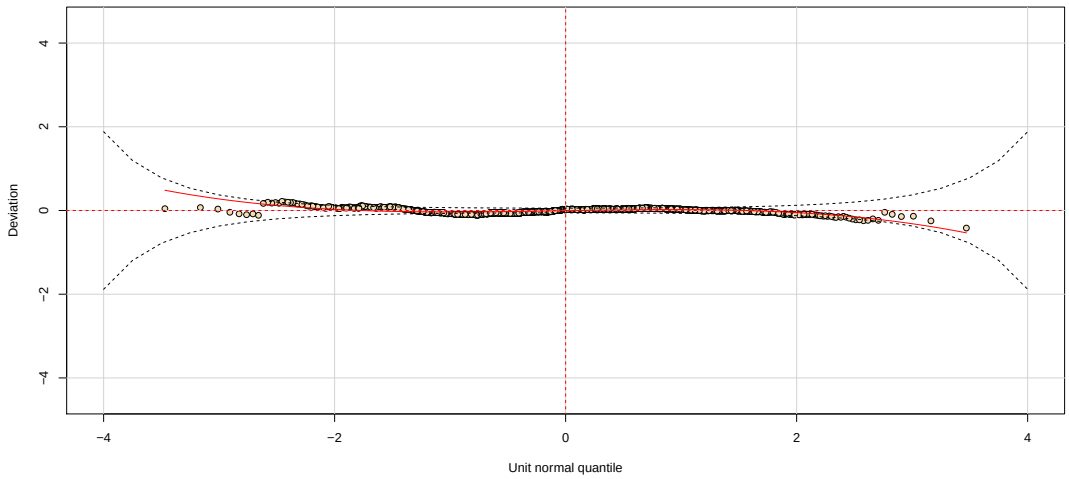

Figura 3.17 – Worm-Plot do Modelo Normal Inverso.

3.1.2.4 Ajuste do Modelo Skew Normal

Dada uma variável aleatória $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$ que segue distribuição SN, o respectivo modelo misto pode ser expresso da seguinte forma:

$$Y_i \sim SN(\mu_{ij}, \sigma^2_{ij}, \lambda_{ij}), \text{ com } i = 1, 2, ..., N \text{ e } j = 1, 2, .., n_i,$$

em que, λ é o parâmetro de assimetria da SN, N é o tamanho da amostra e n_i as respectivas observações repetidas. Os efeitos aleatórios, $b_{i\mu}$ e $b_{i\sigma^2}$ são considerados independentes, seguindo distribuição normal com vetor de médias 0 , matriz de variâncias e covariâncias D_μ e D_{σ^2} e coeficiente de assimetria λ_μ e λ_σ, respectivamente.

De forma análoga, escreve-se:

$$Y_{ij}|b_{i\mu}, b_{i\sigma^2} \sim SN(\mu_{ij}, \sigma^2_{ij}, \lambda_{ij}),$$

em que, $b_{i\mu} \sim N(0, D_{b_\mu})$ e $b_{i\sigma^2} \sim N(0, D_{b^2_\sigma})$.

A distribuição SN, como já visto, possui três parâmetros. Ao utilizar o pacote `gamlss` seria possível modelar os três sem grandes dificuldades, porém como forma de comparação com as demais distribuições estudadas, optou-se em estimar através da modelagem mista apenas μ e σ. Obviamente o parâmetro λ é considerado, mas assume um valor estimado único para o intercepto β_0 e não um valor específico para cada parâmetro como feito para μ e σ.

Contudo, no contexto de GLMM, os parâmetros μ_{ij} e σ^2_{ij}, satisfazem:

$$\begin{aligned} g_\mu &= \eta_{\mu_{ij}} = x^T_{\mu_{ij}}\beta_\mu + Z^T_{\mu_{ij}} b_{i\mu} \\ g_{\sigma^2} &= \eta_{\sigma^2_{ij}} = x^T_{\sigma^2_{ij}}\beta_{\sigma^2} + Z^T_{\sigma^2_{ij}} b_{i\sigma^2}, \end{aligned}$$

onde os componentes $x^T_{\mu_{ij}}, x^T_{\sigma^2_{ij}}$ e $\beta_\mu, \beta_{\sigma^2}$ são referentes à parte fixa do modelo e os componentes $Z^T_{\mu_{ij}}, Z^T_{\sigma^2_{ij}}$ e $b_{i\mu}, b_{i\sigma^2}$ são responsáveis pela parte aleatória.

Definido θ_i como sendo o efeito fixo do $i-$ésimo tratamento (genótipo), α_j o efeito fixo do $j-$ésimo dia de avaliação após a inoculação da bactéria e γ_k o efeito aleatório da $k-$ésima folha destacada. Nessas condições, a Tabela (3.15) mostra os seis modelos que foram ajustados e os respectivos valores de AIC, BIC e GD.

Tabela 3.15 – Modelos ajustados para a distribuição Skew normal.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	3473,83	3490,51	3467,83
2	$\eta_i = \beta_0 + \theta_i trat$	2996,91	3096,99	2960,91
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	826,06	937,26	786,06
4	$\boldsymbol{\eta_{\mathbf{ijk}} = \beta_0 + \theta_{\mathbf{i}}\mathbf{trat} + \alpha_{\mathbf{j}}\mathbf{DAI} + \gamma_{\mathbf{0k}}}$	**805,98**	**1068,91**	**711,40**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	824,06	929,70	786,06
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	803,98	1061,35	711,40

Ao analisar a Tabela (3.15), chega-se as mesmas conclusões obtidas para as demais distribuições quanto a escolha do modelo, ou seja, é importante que o efeito aleatório da folha seja incorporado vistas à notória melhora dos valores dos critérios de seleção, assim como a escolha do modelo quatro como sendo o mais adequado também para a distribuição SN.

Assim, o modelo SN final a ser ajustado considerará a função de ligação logarítmica para σ e μ como já mencionado para os demais ajustes.

As Tabelas (3.16) e (3.17) na sequência apresentam o resumo do modelo que considera como *baseline* a variedade Bahia 25-462 e o DAI 3.

A interpretação das estimativas da Tabela (3.16) possuem resultados semelhantes aos anteriormente obtidos. Observe que o genótipo *Pera 436* continua sendo o de menor crescimento médio das lesões, assim como o genótipo *Morcott 280* que apresenta o segundo menor crescimento médio. A terceira posição é ocupada agora pelo genótipo *Pera 329*. Além disso, os maiores crescimentos médios das lesões são encontrados nos genótipos *Westin 340* e *Rubi 353*.Com base nos DAI, o diâmetro médio das lesões tendem a aumentar com o passar das avaliações, tendo maior diâmetro encontrado no DAI 14.

Tabela 3.16 – Estimativas do parâmetro μ para a distribuição Skew Normal.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	1,363	0,017	< 0,01
Morcott 280	$\boldsymbol{\theta_1}$	**-0,266**	**0,022**	**< 0,01**
Natal 245	θ_2	-0,045	0,050	0,365
Natal 261	θ_3	-0,206	0,027	< 0,01
Natal 308	θ_4	-0,129	0,029	< 0,01
Natal M9-324	θ_5	-0,072	0,015	< 0,01
Natal M9-350	θ_6	-0,189	0,051	< 0,01
Pera 329	θ_7	-0,219	0,029	< 0,01
Pera 331	θ_8	-0,121	0,044	< 0,01
Pera 436	$\boldsymbol{\theta_9}$	**-0,455**	**0,027**	**< 0,01**
Pera 460	θ_{10}	-0,089	0,019	< 0,01
Rubi 251	$\boldsymbol{\theta_{11}}$	**0,044**	**0,093**	**0,634**
Rubi 353	θ_{12}	0,131	0,002	< 0,01
Valência 326	θ_{13}	-0,185	0,023	< 0,01
Westin 16-319	θ_{14}	-0,099	0,194	< 0,01
Westin 340	$\boldsymbol{\theta_{15}}$	**0,219**	**0,098**	**0,025**
DAI 7	α_2	0,377	0,025	< 0,01
DAI 14	α_3	0,801	0,062	< 0,01

A Tabela (3.17) apresenta as estimativas de σ que por si só não possuem interpretação, mas são importantes no contexto geral da modelagem como ja discutido.

Tabela 3.17 – Estimativas do parâmetro σ para a distribuição Skew Normal.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	-2,394	0,107	< 0,01
Morcott 280	θ_1	0,370	0,089	< 0,01
Natal 245	θ_2	0,901	0,089	< 0,01
Natal 261	θ_3	0,455	0,093	< 0,01
Natal 308	θ_4	0,534	0,091	< 0,01
Natal M9-324	θ_5	0,042	0,092	0,648
Natal M9-350	θ_6	0,942	0,091	< 0,01
Pera 329	θ_7	0,512	0,093	< 0,01
Pera 331	θ_8	0,813	0,089	< 0,01
Pera 436	θ_9	0,576	0,104	< 0,01
Pera 460	θ_{10}	0,273	0,091	< 0,01
Rubi 251	θ_{11}	1,402	0,091	< 0,01
Rubi 353	θ_{12}	0,155	0,110	0,159
Valência 326	θ_{13}	0,439	0,097	< 0,01
Westin 16-319	θ_{14}	0,210	0,094	0,025
Westin 340	θ_{15}	1,416	0,091	< 0,01
DAI 7	α_2	0,559	0,065	< 0,01
DAI 14	α_3	1,017	0,062	< 0,01

Além disso, a estimativa do parâmetro λ é apresentada na Tabela (3.18):

Tabela 3.18 – Estimativa do parâmetro λ para a distribuição Skew Normal.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	0,777	0,407	0,056

O diagnóstico gráfico do modelo proposto é dado na sequência, assim como os valores das medidas descritivas de seus resíduos, onde os valores obtidos são satisfatórios ao passo que indicam que estes seguem distribuição normal padrão.

Tabela 3.19 – Medidas descritivas dos resíduos do Modelo Skew Normal.

Modelo Skew Normal	
Média	0,001
Variância	1,018
Coef. Assimetria	0,127
Coef. Curtose	2,809

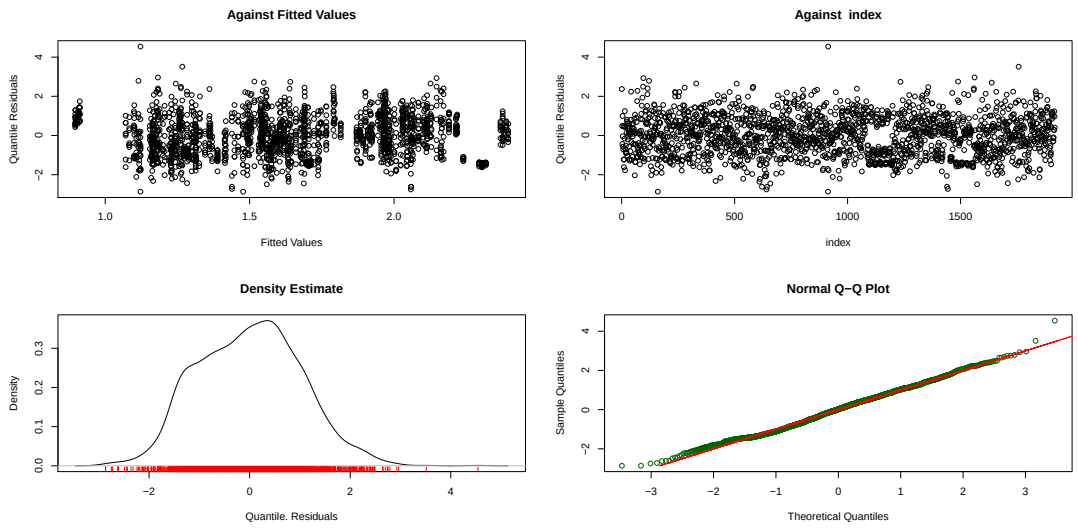

Figura 3.18 – Gráfico de diagnóstico do Modelo Skew Normal.

Finalizando a análise, na Figura (3.19) é apresentado o Worm-Plot, esperando que os valores estejam dentro dos limites de confiança e próximos da linha horizontal em torno de zero.

Assim como para os demais modelos ajustados, a distribuição SN também pode ser considerada uma opção para o ajuste dos dados. Os gráficos utilizados na análise dos resíduos do modelo apresentados nas Figuras (3.18) e (3.19) indicam que os resíduos do modelo SN seguem distribuição aproximadamente normal padrão e que a variabilidade está bem distribuída não assumindo nenhum padrão visível que comprometa a análise.

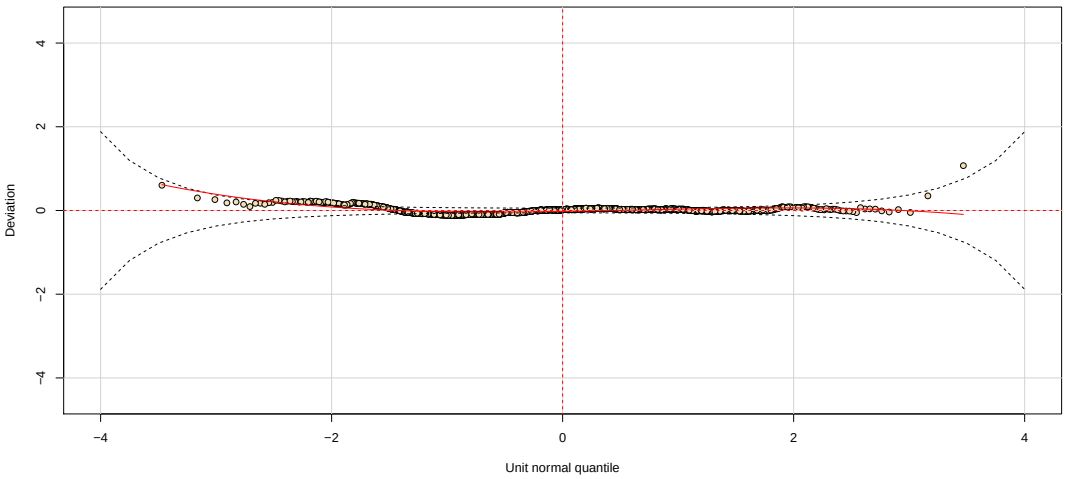

Figura 3.19 – Worm-Plot do Modelo Skew Normal.

Além disso, a Tabela (3.19) evidencia os valores das medidas resumo dos resíduos do modelo, onde estas corroboram os fatos já mencionados anteriormente.

3.1.2.5 Ajuste do Modelo Skew-t tipo 3

Dada uma variável aleatória $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$ que segue distribuição ST3, o respectivo modelo misto pode ser expresso da seguinte forma:

$$Y_i \sim ST3(\mu_{ij}, \sigma^2_{ij}, \nu_{ij}, \tau_{ij}), \text{ com } i = 1, 2, ..., N \text{ e } j = 1, 2, .., n_i,$$

em que, N é o tamanho da amostra e n_i as respectivas observações repetidas. Os efeitos aleatórios, $b_{i\mu}$ e $b_{i\sigma^2}$ são considerados independentes, seguindo distribuição normal com vetor de médias 0 , matriz de variâncias e covariâncias D_μ e D_{σ^2}, respectivamente.

De forma análoga, escreve-se:

$$Y_{ij}|b_{i\mu}, b_{i\sigma^2} \sim ST3(\mu_{ij}, \sigma^2_{ij}, \nu_{ij}, \tau_{ij}),$$

em que, $b_{i\mu} \sim N(0, D_{b_\mu})$ e $b_{i\sigma^2} \sim N(0, D_{b^2_\sigma})$.

A distribuição ST3, como já visto, possui quatro parâmetros. Ao utilizar o pacote *gamlss* seria possível modelar todos eles sem grandes dificuldades, porém como forma de comparação com as demais distribuições estudadas, optou-se em estimar através da modelagem mista apenas μ e σ.

Contudo, no contexto de GLMM, os parâmetros μ_{ij} e σ^2_{ij}, satisfazem:

$$\begin{aligned} g_\mu &= \eta_{\mu_{ij}} = x^T_{\mu_{ij}}\beta_\mu + Z^T_{\mu_{ij}} b_{i\mu} \\ g_{\sigma^2} &= \eta_{\sigma^2_{ij}} = x^T_{\sigma^2_{ij}}\beta_{\sigma^2} + Z^T_{\sigma^2_{ij}} b_{i\sigma^2}, \end{aligned}$$

onde os componentes $x^T_{\mu_{ij}}, x^T_{\sigma^2_{ij}}$ e $\beta_\mu, \beta_{\sigma^2}$ são referentes à parte fixa do modelo e os componentes $Z^T_{\mu_{ij}}, Z^T_{\sigma^2_{ij}}$ e $b_{i\mu}, b_{i\sigma^2}$ são responsáveis pela parte aleatória.

Definido θ_i como sendo o efeito fixo do $i-$ésimo tratamento (genótipo), α_j o efeito fixo do $j-$ésimo dia de avaliação após a inoculação da bactéria e γ_k o efeito aleatório da $k-$ésima folha destacada. Nessas condições, a Tabela (3.20) mostra os seis modelos que foram ajustados e os respectivos valores de AIC, BIC e GD.

Tabela 3.20 – Modelos ajustados para a distribuição Skew-t tipo 3.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	2965,52	2987,76	2957,52
2	$\eta_i = \beta_0 + \theta_i trat$	2941,97	3047,62	2903,97
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	735,91	852,67	693,91
4	$\boldsymbol{\eta_{ijk} = \beta_0 + \theta_i \mathbf{trat} + \alpha_j \mathbf{DAI} + \gamma_{0k}}$	**717,37**	**983,85**	**621,51**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	733,91	845,11	693,91
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	715,37	976,29	621,51

Observando a tabela de modelos ajustados e levando em consideração que comparações entre distribuições sejam feitas, o modelo ST3 escolhido para o ajuste é o modelo de número quatro, assim como para as distribuições anteriores. Além disso, leva-se em consideração que a variável tratamento não é de efeito aleatório no momento em que se encontra a pesquisa, mesmo que testes fazendo seu uso tenham sido feitos.

Dessa forma, o modelo ST3 final a ser ajustado considerará a função de ligação logarítmica para os parâmetros μ e σ. As Tabelas (3.21) e (3.22) na sequência, apresentam essas estimativas:

Tabela 3.21 – Estimativas do parâmetro μ para a distribuição Skew-t tipo 3.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	1,358	0,013	< 0,01
Morcott 280	$\boldsymbol{\theta_1}$	**-0,274**	**0,020**	**< 0,01**
Natal 245	θ_2	-0,038	0,028	0,165
Natal 261	θ_3	-0,201	0,019	< 0,01
Natal 308	θ_4	-0,129	0,021	< 0,01
Natal M9-324	θ_5	-0,086	0,018	< 0,01
Natal M9-350	θ_6	-0,162	0,030	< 0,01
Pera 329	θ_7	-0,218	0,019	< 0,01
Pera 331	θ_8	-0,105	0,026	< 0,01
Pera 436	$\boldsymbol{\theta_9}$	**-0,492**	**0,024**	**< 0,01**
Pera 460	θ_{10}	-0,099	0,018	< 0,01
Rubi 251	$\boldsymbol{\theta_{11}}$	0,072	0,043	0,097
Rubi 353	θ_{12}	**0,113**	**0,018**	**< 0,01**
Valência 326	θ_{13}	-0,195	0,020	< 0,01
Westin 16-319	θ_{14}	-0,104	0,017	< 0,01
Westin 340	$\boldsymbol{\theta_{15}}$	**0,246**	**0,040**	**< 0,01**
DAI 7	α_2	0,398	0,011	< 0,01
DAI 14	α_3	0,814	0,013	< 0,01

Tabela 3.22 – Estimativas do parâmetro σ para a distribuição Skew-t tipo 3.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	-2,497	0,076	< 0,01
Morcott 280	$\boldsymbol{\theta_1}$	0,409	0,095	< 0,01
Natal 245	θ_2	0,879	0,105	<0,01
Natal 261	θ_3	0,431	0,100	< 0,01
Natal 308	θ_4	0,512	0,098	< 0,01
Natal M9-324	θ_5	0,165	0,092	0,074
Natal M9-350	θ_6	0,896	0,100	< 0,01
Pera 329	θ_7	0,479	0,087	< 0,01
Pera 331	θ_8	0,792	0,108	< 0,01
Pera 436	$\boldsymbol{\theta_9}$	0,619	0,148	< 0,01
Pera 460	θ_{10}	0,279	0,107	< 0,01
Rubi 251	$\boldsymbol{\theta_{11}}$	1,375	0,108	<0,01
Rubi 353	θ_{12}	0,224	0,111	0,044
Valência 326	θ_{13}	0,449	0,092	< 0,01
Westin 16-319	θ_{14}	0,201	0,095	0,035
Westin 340	$\boldsymbol{\theta_{15}}$	1,409	0,105	< 0,01
DAI 7	α_2	0,477	0,046	< 0,01
DAI 14	α_3	0,939	0,053	< 0,01

Ao interpretar os valores das estimativas do parâmetro μ, dadas na Tabela (3.21), percebe-se que o genótipo *Pera 436* permanece na primeira posição tendo o menor cresci-

mento médio das lesões observado, seguido pelo genótipo *Morcott 280*. Assim como para a distribuição SN, a terceira posição de menor crescimento médio é ocupada pelo genótipo *Pera 329*, que nas demais análises não havia se mostrado como destaque positivo nem negativo. Quanto aos maiores crescimentos médios observados, tem-se um consenso entre todas as distribuições comparadas nessa aplicação, vistas que os genótipos *Westin 340* e *Rubi 353* continuam sendo os que se destacam.

Observando as estimativas dos parâmetros α_2 e α_3 é notório o crescimento gradual dos diâmetros com o passar dos DAI, assim como já observado nas estimativas dos demais modelos ajustados. Assim, pode-se dizer que as conclusões tiradas a cerca do ajuste ST3 corroboram com a análise descritiva inicial dos dados. Além disso, os resultados se assemelham muito com os obtidos no ajuste SN, fato esperado por termos uma relação entre as distribuições e as caudas mais pesadas da ST3 não influenciarem tanto nos resultados pela amostra utilizada nas análises ser significativamente grande.

Com relação a análise de resíduos do modelo, apresenta-se inicialmente na Tabela (3.23) as medidas descritivas dos resíduos, onde estas indicam um ajuste adequado ao passo que os valores obtidos são próximos aos de referência de uma distribuição normal padrão.

Tabela 3.23 – Medidas descritivas dos resíduos do Modelo Skew-t tipo 3.

Modelo Skew-t tipo 3	
Média	0,029
Variância	0,978
Coef. Assimetria	-0,149
Coef. Curtose	2,748

Por fim, os gráficos de resíduos e worm plot são dados nas Figuras (3.20) e (3.21). Nota-se na Figura (3.20) que a variabilidade dos resíduos está bem distribuída, não assumindo nenhum comportamento adverso e os pontos estão todos dispostos em torno de zero. Quanto ao gráfico q-q plot, os pontos aparentemente estão em sua maioria sob a linha de $45°$, indicando também um ajuste satisfatório dos dados.

A Figura (3.21), apresenta o gráfio de Worm-Plot do modelo, onde espera-se que os pontos estejam todos dentro das bandas de confiança e em torno da linha horizontal de zero. Vale lembrar que a retirada dos pontos discrepantes no início dos ajustes melhorou consideravelmente os gráficos de Worm-Plot, principalmente dos modelos SN e ST3.

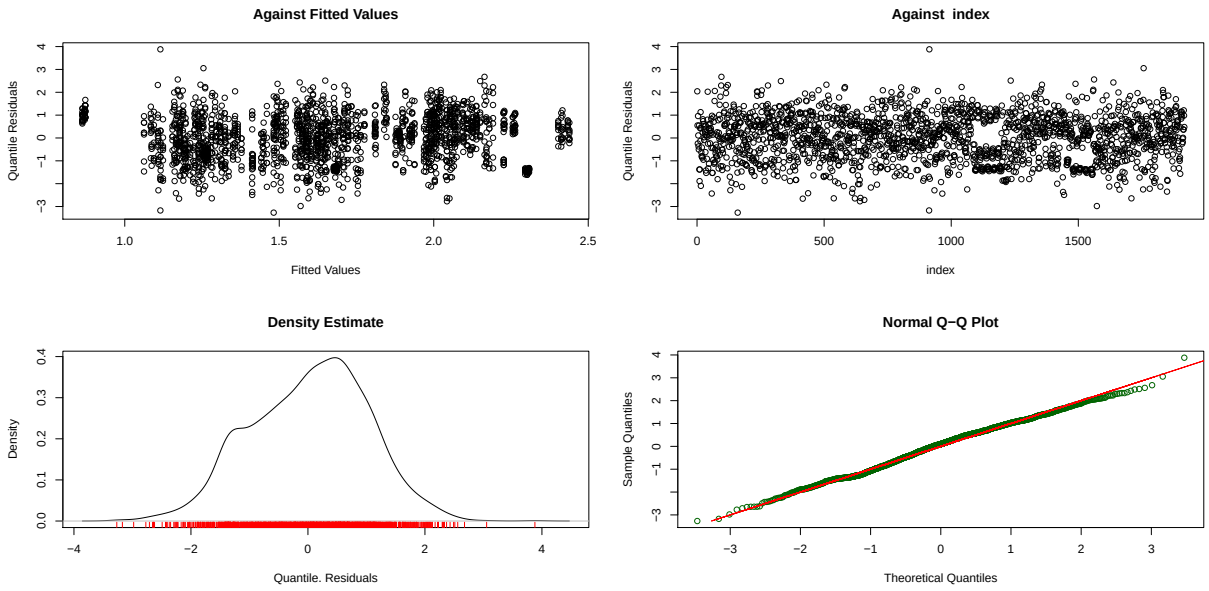

Figura 3.20 – Gráfico de diagnóstico do Modelo Skew-t tipo 3.

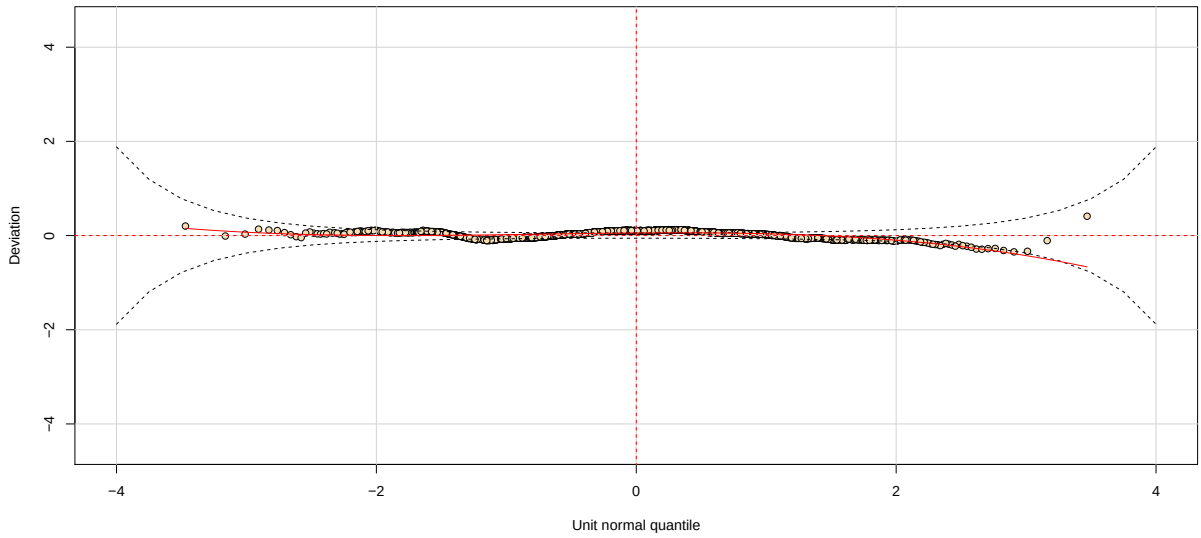

Figura 3.21 – Worm-Plot do Modelo Skew-t tipo 3.

Após a análise dos resíduos ser feita, pode-se observar que tanto os gráficos das Figuras (3.20) e (3.21) quanto a Tabela (3.23), evidenciam que o ajuste é adequado, ao passo que indicam que os resíduos do modelo ST3 seguem possível distribuição normal padrão. Tal fato pode ser observado nas medidas descritivas, assim como nos gráficos de Worm-Plot e q-q plot por exemplo.

3.1.3 Conclusões

De modo geral, as cinco distribuições estudadas apresentaram resultados muito próximos tanto para o modelo escolhido, quanto para as estimativas dos parâmetros. Nesse sentido, os modelos GA, IG e LOGNO possuem forte semelhança nos resultados e se distanciam de forma discreta dos obtidos pelo modelo SN e ST3.

Os modelos corroboraram quanto ao fato do genótipo *Pera 436* ser o que apresenta menor crescimento médio das lesões no decorrer das avaliações, assim como o fato do genótipo *Westin 340* ser o que apresenta o maior crescimento médio. Além disso, os modelos IG, GA e LOGNO "concordam"entre si quanto aos genótipos *Morcott 280* e *Valência 326* seguirem a lista de menor crescimento médio logo após o *Pera 436* e o genótipo *Rubi 251* ocupar o segundo lugar de maior diâmetro médio observado.

Com relação aos modelos SN e ST3, isso não ocorre. A segunda posição de menor diâmetro médio é composta pelo genótipo *Morcott 280*, assim como para os demais modelos, porém a terceira fica a cargo do genótipo *Pera 329* que até então não havia sido mencionado como destaque positivo, nem negativo. Além disso, a segunda posição de maior diâmetro médio agora é ocupada pelo genótipo Rubi 353, que também não havia sido mencionado nas análises anteriores.

Pode-se dizer então que os modelos GA, LOGNO e IG são praticamente correlatos, até mesmo nos valores das estimativas dos parâmetros. Diferindo de forma discreta estão os modelos SN e ST3 que apresentam forte semelhança quando comparados aos demais. Contudo, apesar de algumas discrepâncias, as cinco distribuições ajustadas podem ser consideradas opções para esse tipo de dados.

3.2 Aplicação 2

3.2.1 Análise Descritiva

Como passo inicial, analisou-se o histograma da variável resposta, visto na Figura (3.22). Aparentemente, o comportamento dos dados se difere do estudado até o momento, indicando uma possível simetria para a variável de interesse. Sabendo que graficamente a distribuição normal e t-student se assemelham pela característica de simetria em forma de sino, e se mostram próximas para amostras consideradas grandes, optou-se por avaliar essas duas opções de distribuições para a modelagem.

Essa possível simetria pode ser observada também na Figura (3.22), onde a linha pontilhada representa o gráfico de densidade e a linha contínua as respectivas distribuições de probabilidade t-student e normal ajustadas a partir dos dados:

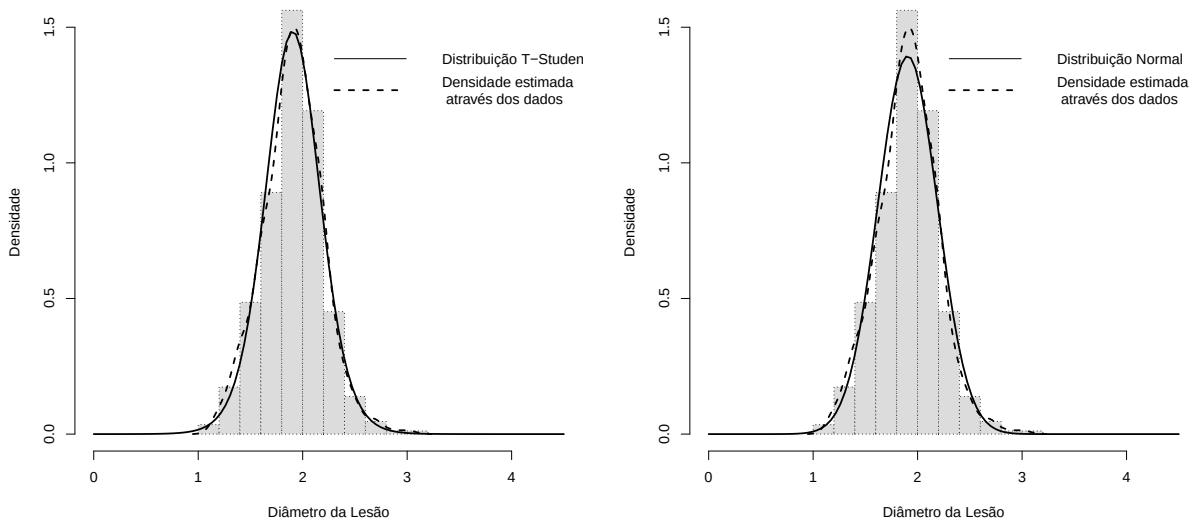

Figura 3.22 – Ajustes das distribuições T-Student e Normal à variável resposta diâmetro da lesão.

A Tabela (3.24) apresenta as medidas resumo em relação a cada um dos seis genótipos de citros estudados.

Observando a Tabela (3.24), verifica-se que os maiores diâmetros médios das lesões são encontrados nos genótipos *Pera Ori* (2,109) e *Prec Ori* (2,102). Os menores diâmetros médios, por sua vez, são vistos nos genótipos *Valência* e *Pera IAC*, assumindo os valores de 1,628 e 1,702, nessa ordem.

Tabela 3.24 – Medidas resumo do diâmetro de lesão para cada variedade.

Genótipos	Min.	Máx.	Média	DP	CV
Hamlin	1,570	2,390	1,986	0,169	0,085
Irradiada	1,490	2,170	1,903	0,136	0,071
Pera IAC	1,190	2,060	1,702	0,229	0,134
Pera Ori	**1,590**	**3,010**	**2,109**	**0,299**	**0,142**
Prec Ori	1,330	2,690	2,102	0,256	0,122
Valência	**1,170**	**2,010**	**1,628**	**0,191**	**0,117**

Com relação ao coeficiente de variação e ao desvio padrão à estes não foi observada a mesma característica. A variabilidade em torno da média é menor para o genótipo *Irradiada* e *Hamlin*, ou seja, seus valores são mais condensados. Os maiores desvios são encontrados nos genótipos *Pera Ori* e *Prec Ori*. A variação média dos diâmetros também é menor para os genótipos *Hamlin* e *Irradiada*. Apesar disso, apresentam-se com os maiores diâmetros mínimos, ou seja, as lesões já são as maiores dentre as menores.

Tais fatos podem ser observados no gráfico de Boxplot apresentado na Figura (3.23):

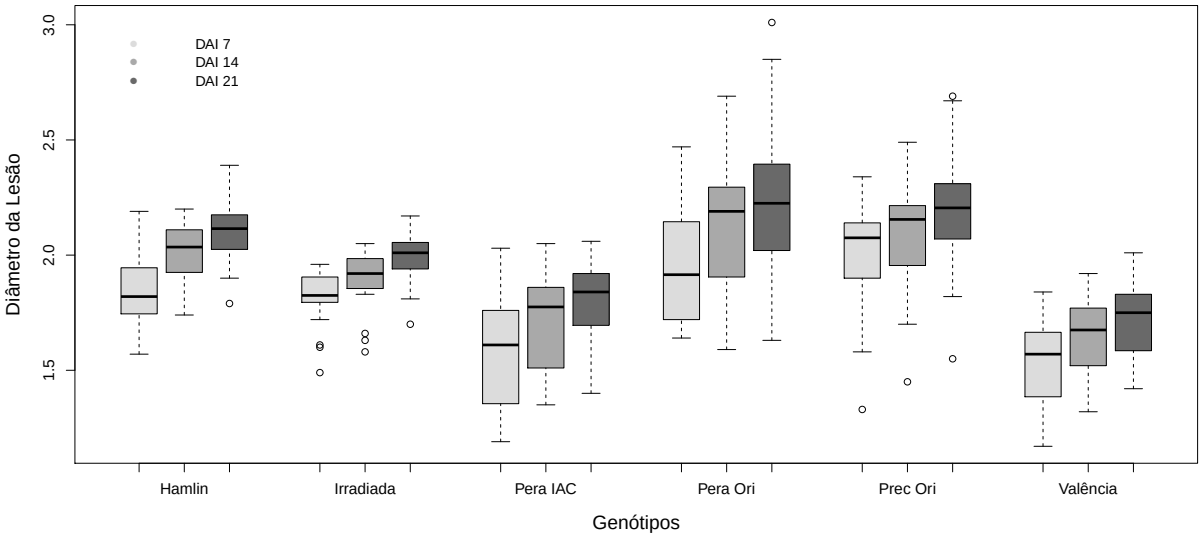

Figura 3.23 – Boxplot da variável diâmetro da lesão para os genótipos em cada DAI.

Corroborando as informações analisadas na tabela de medidas resumo, observa-se na Figura (3.23) que de fato, os menores diâmetros das lesões são encontrados nos genótipos *Valência* e *Pera IAC* em todos os DAI. Assim como os maiores, são vistos em *Pera Ori* e *Prec Ori*. É observável também que os diâmetros estão sempre em crescimento com o passar dos DAI, indicando uma evolução constante da doença.

No que diz respeito aos diâmetros das lesões em cada dia de avaliação após a inocu-

lação da bactéria (DAI), vê-se na Figura (3.24) que nos primeiros sete dias os valores dos diâmetros estão condensados, aparentemente, entre 0 e 2,5 . A partir do décimo quarto dia os diâmetros começam a se distanciar do valor 0, mas ainda assim condensados entre 1 e 2,5. No último dia de avaliação (DAI 21) esses valores se dissipam, ficando a maioria aproximadamente, entre 1,8 e 2,3, mas assumindo valores de 1,5 até 3.

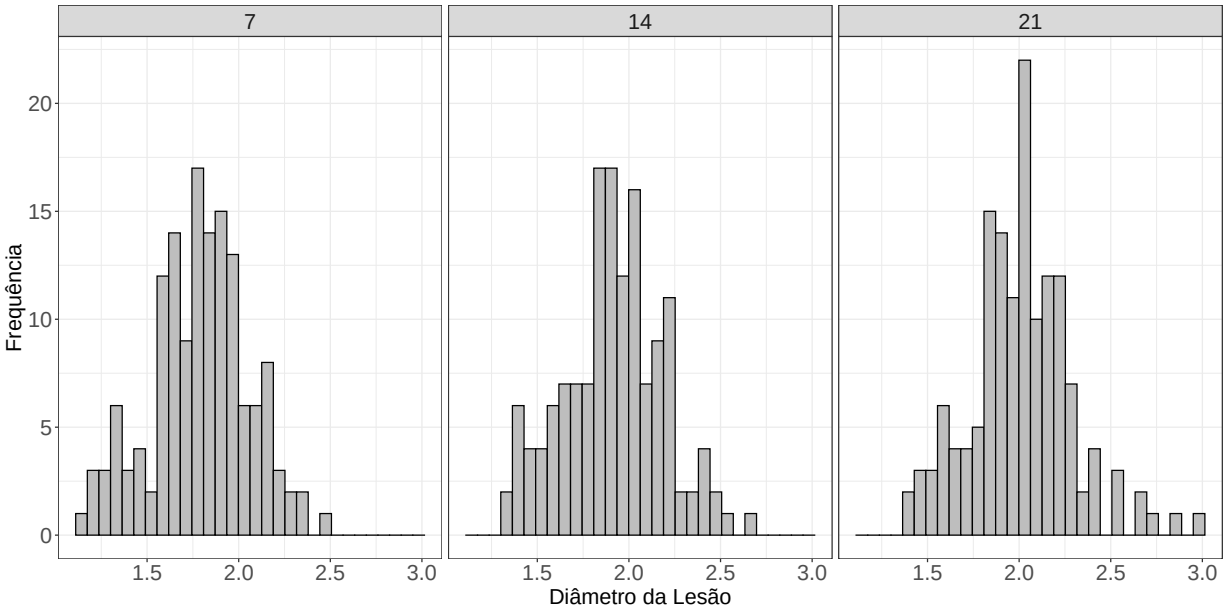

Figura 3.24 – Histograma da variável diâmetro da lesão em cada DAI.

Por fim, sabendo que os gráficos de perfis são ferramentas descritivas importantes na análise de dados longitudinais, a Figura (3.25) relaciona o diâmetro da lesão em cada DAI, levando em consideração os genótipos estudados.

Observando o gráfico de perfis da Figura (3.25), nota-se que de fato o genótipo *Valência* é o que apresenta o menor diâmetro médio das lesões com o passar dos DAI, assim como o maior é encontrado no genótipo *Pera Ori*, corroborando as análises anteriormente feitas através da tabela de medidas resumo e gráfico de boxplot. Percebe-se também que há um crescimento dos diâmetros das lesões, mas que este crescimento não é tão discrepante de uma avaliação para a outra. Fato que pode ser visto nas Figuras (3.24) e (3.23).

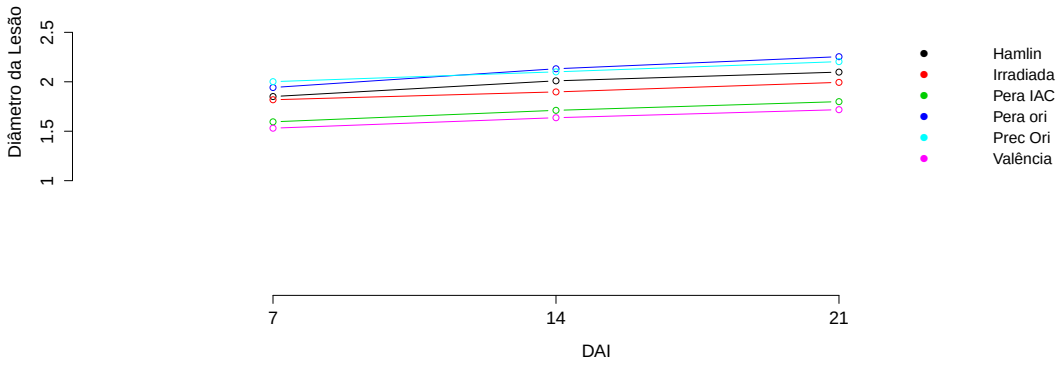

Figura 3.25 – Gráfico de perfis para cada genótipo.

3.2.2 Ajustes

A variável resposta considerada para os modelos é o diâmetro da lesão. As possíveis variáveis explicativas utilizadas são DAI, tratamento e folha, onde as duas primeiras foram consideradas como de efeito fixo e a última como de efeito aleatório. Os modelos foram ajustados para duas distribuições, normal e t-Student considerando a variável resposta y_{ijk}, com $i = 1, ..., 6$ denotando os tratamentos, $j = 1, 2, 3$ denotando os respectivos dias de avaliação dos diâmetros das lesões após a inoculação da bactéria (DAI) e $k = 1, ..., 48$ denotando o número de folhas destacadas utilizadas no experimento.

Após a inspeção inicial do ajuste dos dados, observou-se que as estimativas obtidas pelo modelo normal e t-student foram praticamente idênticas (situação esperada visto que o tamanho amostral para esse caso era razoavelmente grande), se diferindo em alguns casos, apenas na quinta casa decimal. Assim, com intuito de otimizar e simplificar as análises, essas estimativas e demais resultados serão apresentados em conjunto e discutidos em seguida.

3.2.2.1 Ajuste do Modelo Normal e T-Student

A especificação dos modelos normal e t-student seguem as mesmas estruturas tanto para μ quanto para σ dos modelos já especificados na aplicação anterior, utilizando a função de ligação logarítmica.

Considerando então, θ_i o efeito fixo do $i-$ésimo tratamento (genótipo), α_j o efeito fixo do $j-$ésimo dia de avaliação após a inoculação da bactéria e γ_k o efeito aleatório da

$k-$ésima folha destacada. Nessas condições, as Tabela (3.25) e (3.26) mostram os seis modelos que foram ajustados para as distribuições normal e t-student (respectivamente) e os valores de AIC, BIC e GD.

Tabela 3.25 – Modelos ajustados para a distribuição Normal.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	149,044	157,18	145,04
2	$\eta_i = \beta_0 + \theta_i trat$	-73,77	-45,29	-87,77
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	-151,24	-114,62	-169,24
4	$\boldsymbol{\eta_{\mathbf{ijk}} = \beta_0 + \theta_{\mathbf{i}} \mathbf{trat} + \alpha_{\mathbf{j}} \mathbf{DAI} + \gamma_{\mathbf{0k}}}$	**-193,27**	**-132,70**	**-223,04**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	-153,24	-120,69	-169,24
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	-195,27	-138,77	-223,04

Tabela 3.26 – Modelos ajustados para a distribuição T-Student.

	Modelos Ajustados	Critérios de Seleção		
		AIC	BIC	GD
1	$\eta = \beta_0$	147,29	159,49	141,29
2	$\eta_i = \beta_0 + \theta_i trat$	-81,03	-48,49	-97,03
3	$\eta_{ij} = \beta_0 + \theta_i trat + \alpha_j DAI$	-159,90	-117,22	-179,90
4	$\boldsymbol{\eta_{\mathbf{ijk}} = \beta_0 + \theta_{\mathbf{i}} \mathbf{trat} + \alpha_{\mathbf{j}} \mathbf{DAI} + \gamma_{\mathbf{0k}}}$	**-192,81**	**-128,95**	**-224,19**
5	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{1k} trat$	-161,90	-125,29	-179,90
6	$\eta_{ijk} = \beta_0 + \theta_i trat + \alpha_j DAI + \gamma_{0k} + \gamma_{1k} trat$	-194,81	-135,02	-224,19

Avaliando os valores de AIC, GD e BIC nas Tabelas (3.25) e (3.26) dos seis modelos ajustados, é visível que os valores entre as distribuições são muito próximos e que em geral, os efeitos fixos de tratamento e DAI são importantes na análise, ao passo que os valores desses critérios diminuem consideravelmente quando essas variáveis são introduzidas.

O modelo de número quatro inclui o efeito aleatório de folha ao que até então era apenas um modelo de efeitos fixos e evidencia sua importância quando diminui ainda mais os valores desses critérios. Testes com efeito de tratamento devido a escolha de uma folha aleatória para cada genótipo foram feitos nos modelos cinco e seis. Sabe-se que para a finalidade desse trabalho e no momento em que se encontra a pesquisa, tratamento não é uma variável de efeito aleatório, pois não pode ser considerada como uma amostra de uma população maior. Com base nisso, e nos valores apresentados nas duas tabelas, os ajustes são feitos através do modelo de número quatro.

O modelo NO e T final a serem ajustados considerarão a função de ligação logarítmica tanto para o parâmetro μ quanto para σ. Vale lembrar que outras funções de ligação podem e devem ser testadas. Assim, tem-se:

$\log(\mu) = \beta_0$ + θ_1 Irradiada+ θ_2 Pera IAC + θ_3 Pera Ori + θ_4 Prec Ori + θ_5 Valência + θ_6 Hamlin + α_1 DAI 7 + α_2 DAI 14 + α_3 DAI 21 + γ_0 FOLHA 1 + γ_0 FOLHA 2 + ... + γ_0 FOLHA 47 + γ_0 FOLHA 48.

$\log(\sigma) = \beta_0$ + θ_1 Irradiada+ θ_2 Pera IAC + θ_3 Pera Ori + θ_4 Prec Ori + θ_5 Valência + θ_6 Hamlin + α_1 DAI 7 + α_2 DAI 14 + α_3 DAI 21 + γ_0 FOLHA 1 + γ_0 FOLHA 2 + ... + γ_0 FOLHA 47 + γ_0 FOLHA 48.

A Tabela (3.27) mostra as estimativas do parâmetro μ para ambos os modelos NO e T, considerando como *baseline* o genótipo Hamlin e o DAI 7.

Tabela 3.27 – Estimativas do parâmetro μ.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	0,625	0,011	< 0,01
Irradiada	θ_1	-0,022	0,010	0,03
Pera IAC	θ_2	**-0,151**	**0,017**	**<0,01**
Pera Ori	θ_3	0,061	0,019	0,001
Prec Ori	θ_4	0,063	0,016	<0,01
Valência	θ_5	**-0,176**	**0,015**	**<0,01**
DAI 14	α_2	0,051	0,007	<0,01
DAI 21	α_3	0,098	0,007	<0,01

Ao analisar as estimativas dos parâmetros da Tabela (3.27), nota-se que de fato os genótipos *Valência* e *Pera IAC* são os que apresentaram menores crescimentos médios das lesões com o passar das avaliações. Assim como o maior crescimento foi observado para os genótipos *Pera Ori* e *Prec Ori*, corroborando com os resultados apresentados na análise descritiva inicial dos dados. Com relação aos DAI, é evidente (a partir dos valores das estimativas de α_2 e α_3) que o crescimento médio das lesões tende a aumentar com o passar das avaliações.

Apesar de não possuir interpretação prática, ao estimar σ, melhora-se as estimativas de μ. Nesse sentido, as estimativas de σ para as duas distribuições em questão são apresentadas na Tabela (3.28):

Tabela 3.28 – Estimativas do parâmetro σ.

Variável	Parâmetro	Estimativa	Erro Padrão	P-valor
Intercepto	β_0	-1,764	0,096	<0,01
Irradiada	θ_1	-1,087	0,119	<0,01
Pera IAC	θ_2	0,214	0,118	0,070
Pera Ori	θ_3	0,570	0,118	<0,01
Prec Ori	θ_4	0,317	0,118	<0,01
Valência	θ_5	0,038	0,118	0,749
DAI 14	α_2	-0,036	0,084	0,670
DAI 21	α_3	-0,040	0,085	0,634

No que se refere a análise de resíduos dos modelos NO e T, apresentam-se os gráficos de diagnóstico do modelo, bem como o gráfico de Worm-Plot e os valores das medidas descritivas dos resíduos, onde espera-se que esses valores estejam próximos dos valores de referência de uma distribuição normal padrão.

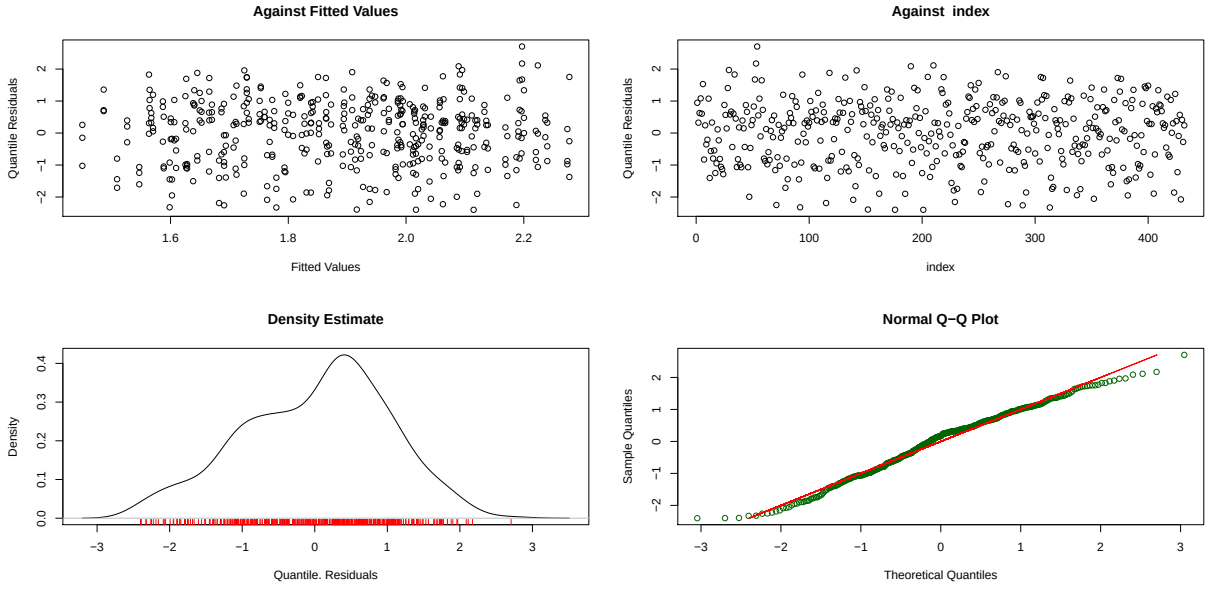

Figura 3.26 – Gráfico de diagnóstico do Modelo Normal.

Analisando as Figuras (3.26) e (3.27) é notória a quase que idêntica semelhança dos gráficos, assim como foi observado nas estimativas dos parâmetros dos dois modelos. Quanto ao comportamento dos resíduos dos modelos, não observa-se nenhum comportamento adverso para a variabilidade, os valores estão todos dispersos em torno da linha horizontal zero, assim como não se observa valores que ultrapassem significativamente o intervalo de -3 a 3. Tais fatos indicam um ajuste adequado.

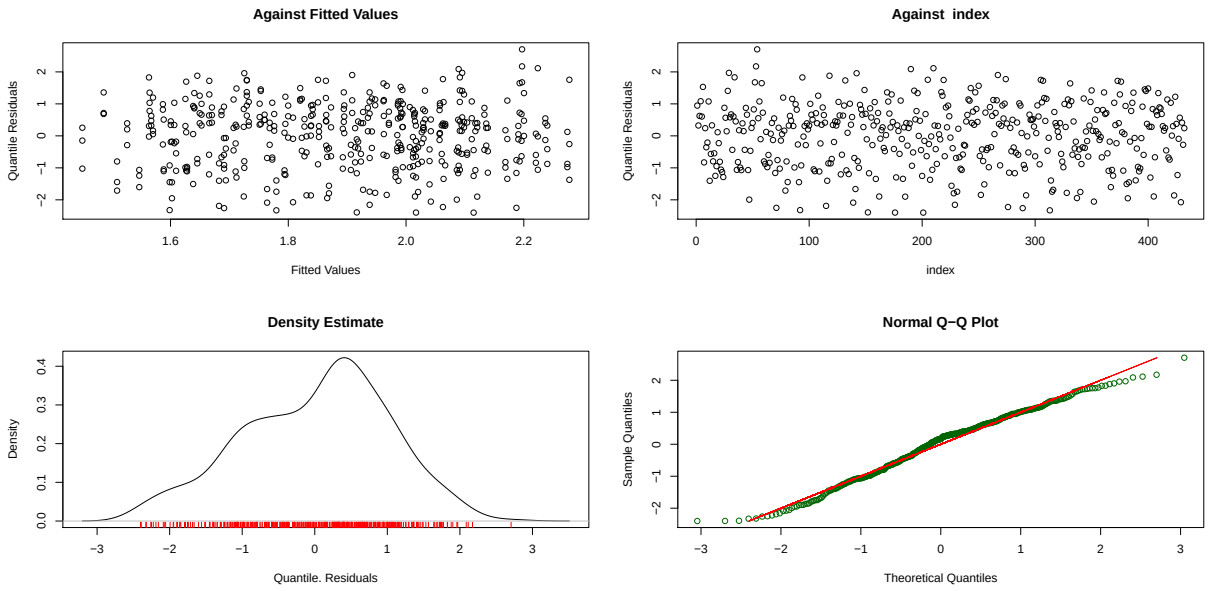

Figura 3.27 – Gráfico de diagnóstico do Modelo T-Student.

Além disso, os gráficos de Worm-Plot dos dois modelos também possuem forte semelhança e portanto será apresentado apenas o do modelo NO na Figura (3.28):

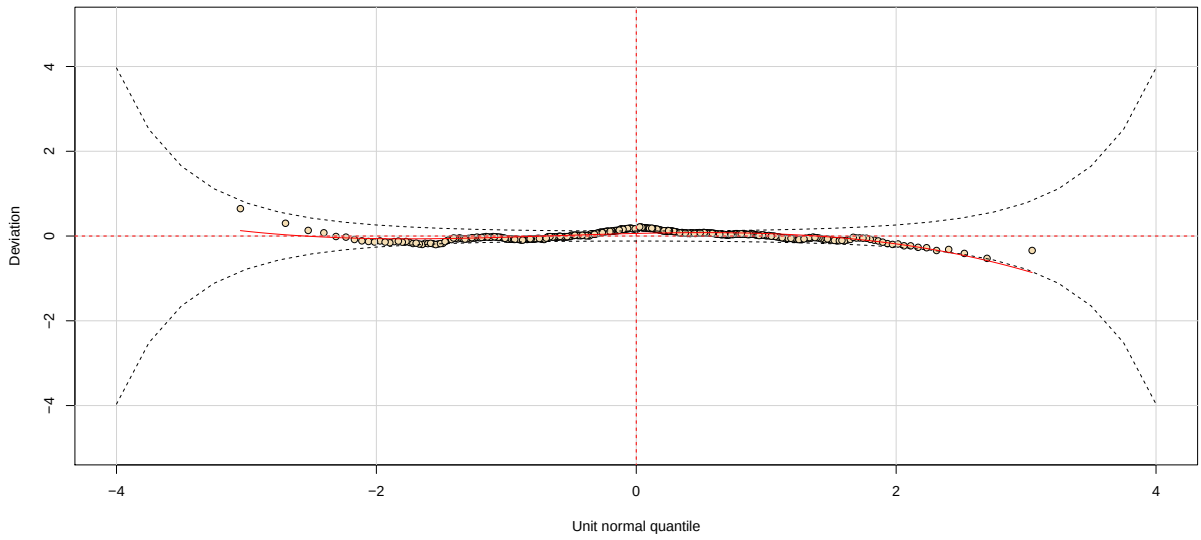

Figura 3.28 – Worm plot do Modelo Normal.

Encerrando a análise diagnóstica dos modelos propostos, as medidas descritivas dos resíduos são dadas na Tabela (3.29), onde indicam que os resíduos possivelmente seguem distribuição normal padrão vistas aos valores da média serem próximos de 0, variância próxima de 1, assimetria próxima de 0 e curtose próxima de 3.

Tabela 3.29 – Medidas descritivas dos resíduos do Modelo Normal e T-Student.

Modelo Normal e T-Student	
Média	0,0154
Variância	1,002
Coef. Assimetria	-0,263
Coef. Curtose	2,557

Assim, relativo a análise diagnóstica dos modelos, pode-se dizer que ambos os modelos apresentaram resíduos satisfatórios, seguindo possível distribuição normal padrão, conforme indicam as medidas descritivas dos resíduos de ambos (apresentadas na Tabela (3.29)) e as Figuras (3.26), (3.27) e (3.28).

3.2.3 Conclusões

Contudo, pode-se dizer que os dois modelos propostos para esse experimento tiveram desempenhos praticamente idênticos tanto nos ajustes, quanto na análise diagnóstica e adequabilidade do modelo aos dados. Por possuírem essa forte semelhança nos resultados, as distribuições normal e t-student são opções quase que indistinguíveis se for considerada uma possível escolha de melhor distribuição a ser utilizada.

3.2.4 Passos Futuros

Este trabalho buscou, em sua essência, encontrar e estudar opções de distribuições de probabilidade que melhor se adequassem aos bancos de dados a fim de utilizá-las nas modelagens e evidenciar quais genótipos de citros seriam mais e menos suscetíveis a doença cancro cítrico. A partir dessa finalidade, estudou-se além das distribuições, algumas metodologias para o ajuste de modelos que conseguissem captar o maior número de informações possíveis dos dados em questão. Contudo, acredita-se que alguns pontos são interessantes de serem explorados e poderiam ser acrescentados ao trabalho, tais como:

- Dedicar estudo a análise de pontos influentes, tipos de perturbação, influência local e glogal;
- Realizar comparações específicas entre as distribuições estudadas em cada um dos experimentos a fim de 'escolher' a 'melhor' em cada um deles.

3.3 Conclusão

O intuito desse trabalho era encontrar um modelo estatístico que fosse capaz de captar o máximo de informações presentes nos dados dos experimentos utilizando a metodologia adequada, bem como ajustar os modelos e fazer a análise diagnóstica. Além disso, um dos principais objetivos das duas aplicações era descobrir dentre os genótipos de citros estudados, quais eram mais/menos suscetíveis ao cancro cítrico.

Como consequência, estudou-se diversas distribuições de probabilidade nos ajustes propostos a fim de encontrar qual delas melhor representasse os conjuntos de dados em ambas as aplicações. Para a primeira delas, comparou-se cinco distribuições de probabilidade que geralmente são utilizadas para modelagem de dados com característica de assimetria positiva. Sendo elas, Gama, Log-normal, Normal Inversa, Skew normal e Skew-t tipo 3.

Apesar de todos os ajustes se apresentarem satisfatórios, algumas comparações mais específicas foram feitas. Os modelos GA, LOGNO e IG se assemelharam muito tanto nos resultados das estimativas quanto nos resíduos. Com relação aos modelos SN e ST3, estes se diferiram de forma discreta dos três anteriormente mencionados, porém compartilharam semelhanças entre si. Essa semelhança pode ser que tenha ocorrido devido ao fato de as distribuições SN e ST3 serem fruto de assimetrizações de duas distribuições que em casos de amostras grandes se aproximam uma da outra.

No que diz respeito aos genótipos estudados na aplicação um, tem-se como resultado que o genótipo *Pera 436* é o que apresenta maior resistência a doença, enquanto o genótipo *Westin 340* mostra-se o mais suscetível.

Na segunda aplicação, o comportamento observado foi de simetria dos dados e assim as distribuições normal e t-student foram as opções consideradas para o ajuste. Em decorrência da forte semelhança nos resultados encontrados nas estimativas dos dois modelos e também nos resíduos, pode-se dizer que as duas distribuições são igualmente adequadas para o banco de dados em questão. No que se refere ao genótipo mais/menos resistente à doença, encontra-se como destaque positivo, ou seja, de maior resistência, o genótipo *Valência* e o de menor *Prec Ori*.

Verificou-se também que a característica de assimetria encontrada nos dados da primeira aplicação não é observada na segunda, ou seja, o comportamento de dados dessa natureza (dados longitudinais com medidas repetidas aplicados na análise da severidade do Cancro Cítrico em citros) não seguem um padrão quanto à possíveis distribuições para serem modelados.

Referências

AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018. 27

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. 29

AKANTZILIOTOU, R. R. K.; STASINOPOULOS, D. The r implementation of generalized additive models for location, scale and shape. *Statistical Modelling Society: Proceedings of the 17th International Workshop on statistical modelling (pp. 75-83)*, 2002. 15, 30

ANSLEY, C. F.; KOHN, R. Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society*, Cambridge University Press, v. 57, n. 3, p. 316–329, 1994. 33

AZZALINI, A. A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, JSTOR, p. 171–178, 1985. 44, 46, 56

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 61, n. 3, p. 579–602, 1999. 44

BARBOSA, M. *Uma abordagem para análise de dados com medidas repetidas utilizando modelos lineares mistos*. Tese (Doutorado) — Tese de Doutorado da Escola Superior de Agricultura Luiz de Queiroz., 2009. 26

BIRNBAUM, Z. W.; SAUNDERS, S. C. A new family of life distributions. *Journal of applied probability*, JSTOR, p. 319–327, 1969. 49

BITANCOURT, A. O cancro cítrico. *O biológico*, v. 23, n. 6, 1957. 20

BOCK, C.; PARKER, P.; GOTTWALD, T. Effect of simulated wind-driven rain on duration and distance of dispersal of xanthomonas axonopodis pv. citri from canker-infected citrus trees. *Plant Disease*, Am Phytopath Society, v. 89, n. 1, p. 71–80, 2005. 20

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, Taylor & Francis Group, v. 88, n. 421, p. 9–25, 1993. 15, 27

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. 36

CHEN, J. T.; GUPTA, A. K.; NGUYEN, T. T. The density of the skew normal sample mean and its applications. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 74, n. 7, p. 487–494, 2004. 47

CNAAN, A.; LAIRD, N. M.; SLASOR, P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, Wiley Online Library, v. 16, n. 20, p. 2349–2380, 1997. 23

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. 32

DESMOND, A. F. On the relationship between two fatigue-life models. *IEEE Transactions on Reliability*, IEEE, v. 35, n. 2, p. 167–169, 1986. 50

DIGGLE, P.; DIGGLE, P. J.; HEAGERTY, P.; LIANG, K.-Y.; HEAGERTY, P. J.; ZEGER, S. et al. *Analysis of longitudinal data*. [S.l.]: Oxford University Press, 2002. 18

DUNCAN, G. J.; KALTON, G. Issues of design and analysis of surveys across time. *International Statistical Review/Revue Internationale de Statistique*, JSTOR, p. 97–117, 1987. 18

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996. 35

FAUSTO, M. A.; CARNEIRO, M.; ANTUNES, C. M. d. F.; PINTO, J. A.; COLOSIMO, E. A. O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. *Cadernos de Saúde Pública*, SciELO Public Health, v. 24, p. 513–524, 2008. 23, 27

FERNÁNDEZ, C.; STEEL, M. F. On bayesian modeling of fat tails and skewness. *Journal of the american statistical association*, Taylor & Francis Group, v. 93, n. 441, p. 359–371, 1998. 47

FERREIRA, W. L.; MORAIES, A. Análise da influência do café no ganho de peso de animais (ratos) por meio de modelo linear misto. *Revista Brasileira de Biometria*, v. 31, p. 485–500, 2013. 25

FOLKS, J. L.; CHHIKARA, R. S. The inverse gaussian distribution and its statistical application—a review. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 40, n. 3, p. 263–275, 1978. 51

FONG, Y.; RUE, H.; WAKEFIELD, J. Bayesian inference for generalized linear mixed models. *Biostatistics*, Oxford University Press, v. 11, n. 3, p. 397–412, 2010. 27

GBUR, E. E.; STROUP, W. W.; MCCARTER, K. S.; DURHAM, S.; YOUNG, L. J.; CHRISTMAN, M.; WEST, M.; KRAMER, M. *Generalized linear mixed models*. [S.l.]: American Society of Agronomy, Crop Science Society of America, Soil Science . . . , 2012. 29

GOLDSTEIN, H. Some models for analysing longitudinal data on educational attainment. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 142, n. 4, p. 407–432, 1979. 17

GONÇALVES-ZULIANI, A. M. O. Resistência de genótipos de laranja doce (citrus sinensis) ao cancro cítrico e diversidade genética de xanthomonas citri subsp. citri. Universidade Estadual de Maringá, 2014. 18

GOTTWALD, T. R.; GRAHAM, J. H.; SCHUBERT, T. S. Citrus canker: the pathogen and its impact. *Plant Health Progress*, Am Phytopath Society, v. 3, n. 1, p. 15, 2002. 20

GRAHAM, J.; GOTTWALD, T. et al. Variation in aggressiveness of xanthomonas campestris pv. citrumelo associated with citrus bacterial spot in florida citrus nurseries. *Phytopathology*, v. 80, n. 2, p. 190–196, 1990. 19

GUEDES, T. A.; ROSSI, R. M.; MARTINS, A. B. T.; JANEIRO, V.; CARNEIRO, J. W. P. Applying regression models with skew-normal errors to the height of bedding plants of stevia rebaudiana (bert) bertoni. *Acta Scientiarum. Technology*, Universidade Estadual de Maringa, v. 36, n. 3, p. 463–468, 2014. 47

HARVILLE, D. A.; MEE, R. W. A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, JSTOR, p. 393–408, 1984. 24

HEDEKER, D.; GIBBONS, R. D. *Longitudinal data analysis*. [S.l.]: John Wiley & Sons, 2006. v. 451. 17

HENDERSON, C. R. Estimation of changes in herd environment. *J. Dairy Sci*, v. 32, n. 8, p. 706–706, 1949. 24

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, JSTOR, p. 423–447, 1975. 24

JR, R. L.; MOHAN, S. Integrated management of the citrus bacterial canker disease caused by xanthomonas campestris pv. citri in the state of paraná, brazil. *Crop Protection*, Elsevier, v. 9, n. 1, p. 3–7, 1990. 20

KHAN, N.; ALI, A.; AHMAD, M.; NOUMAN, M.; ISLAM, B. Evaluation and screening of sweet orange cultivars for vegetative growth and citrus canker. *Sarhad Journal of Agriculture*, v. 32, n. 2, p. 121–126, 2016. 19

KOIZUMI, M. Citrus canker: The world situation. *Citrus Canker: An International Perspective. LW Timmer, ed. University of Florida, Lake Alfred*, p. 2–7, 1985. 20

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. *Biometrics*, JSTOR, p. 963–974, 1982. 14, 17, 24, 25, 26

LEE, K.; LEE, J.; HAGAN, J.; YOO, J. K. Modeling the random effects covariance matrix for generalized linear mixed models. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 6, p. 1545–1551, 2012. 27

LITTELL, R. C.; HENRY, P.; AMMERMAN, C. B. Statistical analysis of repeated measures data using sas procedures. *Journal of animal science*, Oxford University Press, v. 76, n. 4, p. 1216–1231, 1998. 14, 18

MANCO, O. C. U. *Modelos de regressao beta com efeitos aleatórios normais enao normais para dados longitudinais*. Tese (Doutorado) — Tese de doutorado do programa de pós-graduaçao do instituto de matemática e . . . , 2013. 25

MCCULLAGH, P.; NELDER, J. A. Generalized linear models 2nd edition chapman and hall. *London, UK*, 1989. 38, 56

MICHEL, L.; BRUN, F.; MAKOWSKI, D. A framework based on generalised linear mixed models for analysing pest and disease surveys. *Crop Protection*, Elsevier, v. 94, p. 1–12, 2017. 27

NANAMI, D. S. Y. Avaliação de genótipos de laranja doce (citrus sinensis) á xanthomonas citri subsp. citri. 2017. 18

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. 15, 27

OLIVEIRA, V. R. de; RESENDE, M. D. V. de; NASCIMENTO, C. d. S.; DRUMOND, M. A.; SANTOS, C. A. F. Variabilidade genética de procedências e progênies de umbuzeiro via metodologia de modelos lineares mistos (reml/blup). *Embrapa Semiárido-Artigo em periódico indexado (ALICE)*, SciELO Brasil, 2004. 25

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos aditivos generalizados para posição, escala e forma (gamlss) na modelagem de curvas de referência. *Rev. bras. ciênc. saúde*, v. 12, n. 3, p. 289–310, 2008. 35

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. 26

QUEIROZ, M. M. de. Família de distribuições log-skew-multivariadas: definição, entropia e outras propriedades. Universidade Federal de Minas Gerais, 2013. 47

RIECK, J. R.; NEDELMAN, J. R. A log-linear model for the birnbaum—saunders distribution. *Technometrics*, Taylor & Francis, v. 33, n. 1, p. 51–60, 1991. 50

RIGBY, R.; STASINOPOULOS, D. The gamlss project: a flexible approach to statistical modelling. *Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, p. 249–256, 2001. 15, 30

RIGBY, R. A.; STASINOPOULOS, D. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, Springer, v. 6, n. 1, p. 57–65, 1996. 32

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. 15, 30, 31, 32, 34

RIGBY, R. A.; STASINOPOULOS, M. D.; HELLER, G. Z.; BASTIANI, F. D. *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. [S.l.]: CRC press, 2019. 32

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. 29

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. *Variance components*. [S.l.]: John Wiley & Sons, 2009. v. 391. 26

SINGER, J. M.; NOBRE, J. S.; ROCHA, F. M. M. Análise de dados longitudinais apêndices a, b e cversao parcial. 2018. 61

STASINOPOULOS, D. M.; RIGBY, R. A. et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, v. 23, n. 7, p. 1–46, 2007. 30, 33, 34

STASINOPOULOS, M.; RIGBY, B.; AKANTZILIOTOU, C. *Instructions on how to use the gamlss package in R Second Edition*. 2008. 36, 40, 41, 47, 48, 54

TIMMER, L.; GARNSEY, S.; BROADBENT, P. Diseases of citrus. *Diseases of Tropical Fruits Crops. CABI Publishing, CABI International, Wallingford, UK*, p. 163–195, 2003. 19

TWEEDIE, M. C. Inverse statistical variates. *Nature*, Nature Publishing Group, v. 155, n. 3937, p. 453–453, 1945. 50

TWEEDIE, M. C. et al. Statistical properties of inverse gaussian distributions. i. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 362–377, 1957. 50

TWEEDIE, M. C. et al. Statistical properties of inverse gaussian distributions. ii. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 28, n. 3, p. 696–705, 1957. 50

VERBEKE, G.; LESAFFRE, E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, Taylor & Francis, v. 91, n. 433, p. 217–221, 1996. 24

WARE, J. H. Linear models for the analysis of longitudinal studies. *The American Statistician*, Taylor & Francis Group, v. 39, n. 2, p. 95–101, 1985. 17, 18

WILLETT, J. B.; SINGER, J. D.; MARTIN, N. C. The design and analysis of longitudinal studies of development and psychopathology. *Development and psychopathology*, v. 10, p. 395–426, 1998. 18

XAVIER, L. H. *Modelos univariado e multivariado para análise de medidas repetidas e verificação da acurácia do modelo univariado por meio de simulação*. Tese (Doutorado) — Universidade de São Paulo, 2000. 29, 106

YOON, S.; JAIN, A. K. Longitudinal study of fingerprint recognition. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 112, n. 28, p. 8555–8560, 2015. 18

ZHAO, Y.; STAUDENMAYER, J.; COULL, B. A.; WAND, M. P. General design bayesian generalized linear mixed models. *Statistical science*, JSTOR, p. 35–51, 2006. 27

Anexos

ANEXO A

Estruturas das Matrizes de Variâncias e Covariâncias

De acordo com Xavier (2000), e considerando $n_i = 4$, algumas estruturas possíveis para as matrizes de variâncias e covariâncias são apresentadas a seguir:

- Não Estruturada (UN): Assume-se correlações independentes, onde todas são calculadas a partir dos dados.

$$V_i = \begin{pmatrix} \sigma_{11} & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_{22} & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix}$$

- Auto-regressiva de Primeira Ordem (AR(1)): Utilizada em dados igualmente espaçados de séries temporais, onde a covariância entre duas observações decresce a medida que o intervalo entre elas cresce. O parâmetro auto-regressivo é ρ e para um processo estacionário, assume $|\rho| < 1$.

$$V_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

- Auto-regressiva de Primeira Ordem Heterogênea (ARH(1)): Mais comumente utilizada para dados de séries temporais, onde as variâncias e covariâncias são desiguais. O parâmetro auto-regressivo é ρ, satisfazendo $|\rho| < 1$.

$$V_i = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$$

- Auto-regressiva de Primeira Ordem Médias Móveis (ARMA(1,1)): Utilizada em dados de séries temporais,com parâmetro auto-regressivo ρ, componente de médias móveis γ.

$$V_i = \sigma^2 \begin{pmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^1 & \gamma\rho & \gamma & 1 \end{pmatrix}$$

- Simetria Composta (CS): Supõe igualdade de variâncias e covariâncias, ou seja, covariâncias constantes entre as observações de uma mesma unidade amostral devido a erros independentes.

$$V_i = \begin{pmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2 + \sigma_1^2 \end{pmatrix}$$

- Toeplitz (TOEP): Utilizada para dados igualmente espaçados de séries temporais, com correlação arbitrária para cada defasagem.

$$V_i = \begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$$