

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”  
FACULDADE DE CIÊNCIAS AGRONÔMICAS  
CAMPUS DE BOTUCATU

**METODOLOGIAS ESTATÍSTICAS NA ANÁLISE DE  
GERMINAÇÃO DE SEMENTES DE MAMONA**

**LUCIANO BARBOSA**

Tese apresentada à Faculdade de Ciências  
Agronômicas da UNESP – Campus de  
Botucatu, para obtenção do título de Doutor  
em Agronomia (Energia na Agricultura).

BOTUCATU – SP

Outubro – 2010

UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”  
FACULDADE DE CIÊNCIAS AGRONÔMICAS  
CAMPUS DE BOTUCATU

**METODOLOGIAS ESTATÍSTICAS NA ANÁLISE DE  
GERMINAÇÃO DE SEMENTES DE MAMONA**

**LUCIANO BARBOSA**

**Orientadora: Prof<sup>ª</sup>. Dr.<sup>a</sup>. LUZIA APARECIDA TRINCA**

Tese apresentada à Faculdade de Ciências  
Agronômicas da UNESP – Campus de  
Botucatu, para obtenção do título de Doutor  
em Agronomia (Energia na Agricultura).

BOTUCATU – SP

Outubro – 2010

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉCNICA DE AQUISIÇÃO E TRATAMENTO DA INFORMAÇÃO - SERVIÇO TÉCNICO DE BIBLIOTECA E DOCUMENTAÇÃO - UNESP - FCA - LAGEADO - BOTUCATU (SP)

Barbosa, Luciano, 1971-  
B238m      Metodologias estatísticas na análise de germinação de sementes de mamona / Luciano Barbosa. - Botucatu : [s.n.], 2010  
            ix, 108 f. : gráfs., tabs., fots. color.

      Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrônômicas, Botucatu, 2010  
      Orientador: Luzia Aparecida Trinca  
      Inclui bibliografia.

      1. Mamona. 2. Equação de estimação generalizada. 3. Modelo linear generalizado misto. I. Trinca, Luzia Aparecida. II. Universidade Estadual Paulista "Júlio de Mesquita Filho" (Campus de Botucatu). Faculdade de Ciências Agrônômicas. III. Título.

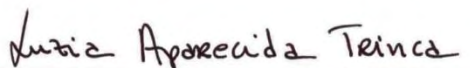
**UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"**  
**FACULDADE DE CIÊNCIAS AGRONÔMICAS**  
**CAMPUS DE BOTUCATU**  
**CERTIFICADO DE APROVAÇÃO**

TÍTULO: "METODOLOGIAS ESTATÍSTICAS NA ANÁLISE DE GERMINAÇÃO DE  
SEMENTES DE MAMONA"

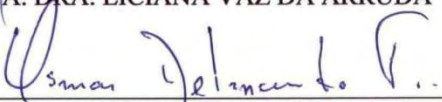
ALUNO: LUCIANO BARBOSA

ORIENTADORA: PROFA. DRA. LUZIA APARECIDA TRINCA

Aprovado pela Comissão Examinadora

  
PROFA. DRA. LUZIA APARECIDA TRINCA

  
PROFA. DRA. LICIANA VAZ DA ARRUDA

  
PROF. DR. OSMAR DELMANTO JUNIOR

  
PROFA. DRA. CÉLIA REGINA LOPES ZIMBACK

  
PROFA. DRA. MARLI TEIXEIRA DE A. MINHONI

Data da Realização: 16 de novembro de 2010.

## **AGRADECIMENTOS**

À Prof<sup>a</sup>. Dr<sup>a</sup>. Luzia Aparecida Trinca, pela orientação, sugestões e paciência durante a realização deste trabalho.

À Prof<sup>a</sup>. Dr<sup>a</sup>. Liciane Vaz de Arruda Silveira, pelo apoio e disposição para discutir assuntos referentes a este trabalho.

À Engenheira Agrônoma Líbia Belisário da Silva por ter cedido o conjunto de dados utilizado neste trabalho.

À Faculdade de Ciências Agronômicas de Botucatu, pela oportunidade de realização do Curso de Doutorado em Agronomia – Energia na Agricultura, fundamental para meu crescimento pessoal e profissional.

Aos professores, funcionários, colegas de pós-graduação e a todos aqueles que, de alguma forma, contribuíram pela realização deste trabalho.

À minha esposa Rose e ao meu filho Thales, pelo carinho, paciência e incentivo.

## SUMÁRIO

	Página
LISTA DE TABELAS .....	VII
LISTA DE FIGURAS .....	VIII
1 RESUMO .....	1
2 SUMMARY .....	2
3 INTRODUÇÃO .....	3
4 REVISÃO BIBLIOGRÁFICA .....	6
4.1 O Biodiesel .....	6
4.2 Caracterização da mamoneira .....	8
4.3 Análise de Sobrevivência .....	14
4.3.1 Estimção não-paramétrica.....	18
4.3.1.1 Estimador de Kaplan-Meier.....	19
4.3.2 Modelos para tempo de vida discreto.....	20
4.3.2.1 Modelo de riscos proporcionais de Cox .....	21
4.3.2.2 Modelo Logístico .....	23
4.4 Modelo Linear Generalizado .....	25
4.4.1 Quase-verossimilhança.....	31
4.4.2 Qualidade do ajuste e seleção de modelos.....	32
4.4.3 Análise de resíduos e diagnósticos .....	34
4.5 O Método da Equação de Estimção Generalizada .....	36
4.6 Modelo Linear Generalizado Misto .....	44
4.6.1 Estimção dos parâmetros por máxima verossimilhança .....	47
5 MATERIAL E MÉTODOS .....	48
5.1 Material .....	48
5.2 Métodos .....	49

6 RESULTADOS E DISCUSSÃO.....	56
6.1 Análise descritiva dos dados pelo método de Kaplan-Meier.....	56
6.2 Análise considerando a proporção de sementes germinadas no final do período.....	59
6.3 Análise considerando a proporção de sementes germinadas por período.....	73
7 CONCLUSÕES.....	93
8 REFERÊNCIAS BIBLIOGRÁFICAS .....	95
APÊNDICE A: O Método de Newton-Raphson .....	102
APÊNDICE B: Comandos no SAS e no R.....	104

## LISTA DE TABELAS

<b>Tabela</b>	<b>Página</b>
1. Componentes da família exponencial para algumas distribuições conhecidas.....	26
2. <i>Deviances</i> para algumas distribuições conhecidas.....	33
3. Épocas de colheita dos racemos de mamoneira .....	49
4. Estimativas da função de sobrevivência pelo método de Kaplan-Meier.....	57
5. Estimativas da função de sobrevivência por Kaplan-Meier.....	59
6. Resultados do ajuste do MLG1 e MLG2.....	60
7. Critérios para verificação do ajuste do MLG1 e MLG2.....	63
8. Resultados do ajuste do MLGM1 e MLGM2.....	64
9. Critério para verificação do ajuste do MLGM1 e MLGM2.....	67
10. Resultados do ajuste do EEG1 e EEG2.....	68
11. Critério para verificação do ajuste do EEG1 e EEG2.....	71
12. Resultados do ajuste do MLG3 e MLG4.....	73
13. Critérios para verificação do ajuste do MLG3 e MLG4.....	77
14. Resultados do ajuste do MLGM3 e MLGM4.....	78
15. Critério para verificação do ajuste do MLGM3 e MLGM4.....	83
16. Resultados do ajuste do EEG3 e EEG4.....	83
17. Critério para verificação do ajuste do EEG3 e EEG4.....	88



## LISTA DE FIGURAS

<b>Figura</b>	<b>Página</b>
1. Plantação de mamona.....	11
2. Censura do tipo I.....	14
3. Censura do tipo II.....	15
4. Censura do tipo aleatório.....	15
5. Censura intervalar.....	16
6. Gráficos Kaplan-Meier para cada tratamento, de acordo com as épocas de colheita.....	58
7. Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG1.....	61
8. Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG2.....	62
9. Gráfico de probabilidade normal para os resíduos de Pearson do MLG1.....	62
10. Gráfico de probabilidade normal para os resíduos de Pearson do MLG2.....	63
11. Gráfico dos resíduos de Pearson condicional versus a ordem das observações para o MLGM1.....	65
12. Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM2.....	66
13. Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM1.....	66
14. Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM2.....	67
15. Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG1..	69
16. Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG2..	70
17. Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG1.....	70
18. Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG2.....	71
19. Gráfico das estimativas da probabilidade de germinação das sementes (MLGM2).....	72
20. Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG3.....	75

21. Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG4.....	76
22. Gráfico de probabilidade normal para os resíduos de Pearson do MLG3.....	76
23. Gráfico de probabilidade normal para os resíduos de Pearson do MLG4.....	77
24. Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM3.....	81
25. Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM4.....	81
26. Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM3.....	82
27. Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM4.....	82
28. Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG3..	86
29. Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG4..	86
30. Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG3.....	87
31. Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG4.....	87
32. Gráfico das estimativas da probabilidade de germinação das sementes no período 7 (MLGM4).....	88
33. Gráfico das estimativas da probabilidade de germinação das sementes no período 14 (MLGM4).....	89
34. Gráfico das estimativas da probabilidade de germinação das sementes no período 21 (MLGM4).....	89
35. Curvas de germinação para as épocas 44, 58, 72 e 86 (MLGM4).....	90
36. Curvas de germinação para as épocas 100, 114, 128 e 142 (MLGM4).....	91

## 1 RESUMO

É bastante comum na área agrícola, experimentos cujas variáveis respostas são contagens ou proporções. Para esse tipo de dados, utiliza-se a metodologia de modelos lineares generalizados quando as respostas são independentes. Por outro lado, quando as respostas são dependentes, há uma correlação entre as observações e isso tem que ser levado em consideração na análise, para evitar inferências incorretas sobre os coeficientes de regressão. Na literatura há técnicas disponíveis para a modelagem e análise desses dados, sendo os modelos disponíveis extensões dos modelos lineares generalizados. No presente trabalho, utiliza-se a metodologia de equação de estimação generalizada, que inclui no modelo uma matriz de correlação para a obtenção de um melhor ajuste. Outra alternativa, também abordada neste trabalho, é a utilização de um modelo linear generalizado misto, no qual o uso de efeitos aleatórios também introduz uma correlação entre observações que tenham algum efeito em comum. Essas duas metodologias são aplicadas a um conjunto de dados obtidos de um experimento para avaliar certas condições na germinação de sementes de mamona da cultivar AL Guarany 2002, com o objetivo de se verificar qual o melhor modelo de estimação para esses dados.

---

**Palavras-chave:** Modelo linear generalizado misto; equação de estimação generalizada.

**STATISTICAL METHODS FOR THE ANALYSIS OF SEED GERMINATION OF CASTOR BEAN. Botucatu, 2010. 117p. Tese (Doutorado em Agronomia/Energia na Agricultura) - Faculdade de Ciências Agrônômicas, Universidade Estadual Paulista.**

Author: LUCIANO BARBOSA

Adviser: Ph. D. LUZIA APARECIDA TRINCA

## **2 SUMMARY**

Experiments whose response variables are counts or proportions are very common in agriculture. For this type of data, if the observational units are independent, the methodology of generalized linear models can be appropriate. On the other hand, when responses are dependent or clustered, there is a correlation between the observations and that has to be taken into consideration in the analysis to avoid incorrect inferences about the regression coefficients. In the literature there are techniques available for modeling and analyzing such type data, the models being extensions of generalized linear models. The present study explores the use of: 1) generalized estimation equations, that includes a correlation matrix to obtain a better fit; 2) generalized linear mixed models, that introduce a correlation between clustered observations through the addition of random effects in the model. These two methodologies are applied to a data set obtained from an experiment to evaluate certain conditions on the germination of seeds of castor bean cultivar AL Guarany 2002 with the objective of determining the best estimation model for such data.

---

**Keywords:** Generalized linear mixed model; generalized estimation equation.

### **3 INTRODUÇÃO**

Atualmente, devido aos elevados custos do preço do barril de petróleo, os combustíveis alternativos passaram a ser competitivos do ponto de vista econômico e promissores para o futuro próximo. As reservas de petróleo tendem à escassez e o desenvolvimento mundial dependerá da tecnologia e infraestrutura para utilização de fontes alternativas de energia.

Uma das opções que se tem estudado é a substituição do diesel pelo biodiesel, que pode ser obtido por fontes renováveis como óleos vegetais (amendoim, algodão, soja, mamona e diversas outras plantas oleaginosas) ou gorduras animais. O biodiesel possui um forte apelo social, por se tratar de uma fonte de energia renovável e limpa. É capaz de gerar emprego e renda no campo, assim como diminuir a dependência brasileira dos mercados internacionais de energia.

Comparado ao óleo diesel derivado do petróleo, o biodiesel pode reduzir em 78% as emissões líquidas de gás carbônico, considerando-se a reabsorção pelas plantas. Além disso, reduz em 90% as emissões de fumaça e praticamente elimina as emissões de óxido de enxofre (HOLANDA, 2004).

O Brasil apresenta grande potencial na produção de biodiesel, destacando a utilização de fatores de produção baratos e abundantes, como disponibilidade de mão de obra e terra. O país possui a maior extensão de terras não utilizada para produção

agrícola, além de outras em degradação. Há aproximadamente 90 milhões de hectares de terras que podem ser utilizadas no processo produtivo. Na região do cerrado, por exemplo, mais de 20 milhões de hectares poderão ser utilizados nos próximos anos na integração agricultura-pastagem.

Atualmente, existe uma grande discussão sobre a produção de biodiesel a partir da utilização de óleo de mamona, principalmente para a região do semiárido nordestino. Nessa região, concentra-se a produção nacional de mamona com cerca de 80% da produção total (HOLANDA, 2004). A cultura de mamona é importante para a economia dessa região, por ser resistente à seca, fixadora de mão de obra, geradora de empregos e de matéria-prima para diversos usos.

O óleo de mamona é adequado para produzir biodiesel, por ser solúvel em álcool e não necessitar de calor para a sua transformação para combustível. Do óleo de mamona também são obtidos cosméticos, lubrificantes, plastificantes, detergentes, graxas, tintas, vernizes, poliuretanos, próteses humanas, defensivos agrícolas, entre outros.

O fato do Brasil ser atualmente o terceiro produtor mundial de mamona levou o Governo Federal a eleger o uso da cultura de mamona como forma de promover a inclusão social, aumentando a demanda por sementes de mamona de qualidade, fazendo-se necessário mais estudos dessa oleaginosa, cujo cultivo não é exclusivo do semiárido, sendo também plantada com excelentes resultados em diversas regiões do país.

Vários estudos têm sido realizados para verificar a viabilidade da mamona na produção do biodiesel. Metodologias estatísticas são necessárias para a análise desses dados. Se a variável de interesse é o tempo até a ocorrência de um determinado evento, como por exemplo, tempo de vida de plantas sujeitas a diferentes tratamentos ou tempo até a germinação das sementes, temos que usar uma metodologia que leve em conta a presença de censura, a qual é a observação parcial da resposta. Além disso, esses dados geralmente são observados semanalmente, o que não nos permite saber o tempo exato de ocorrência do evento, e sim, do intervalo em que ele ocorreu. Dados desse tipo apresentam um grande número de empates e são conhecidos como dados de sobrevivência agrupados.

O conjunto de técnicas estatísticas para análise de dados com censura é denominado Análise de Sobrevida, e o intervalo de tempo até a ocorrência do evento é chamado de tempo de sobrevivência, tempo de vida ou tempo de falha.

Outro problema que surge é quando várias sementes são colocadas num mesmo rolo de papel toalha tipo Germitest para avaliar o potencial germinativo. As respostas obtidas nesse rolo de papel são correlacionadas, ou seja, devem ser modeladas através da inclusão de uma matriz de variâncias-covariâncias não diagonal. Isso pode ser feito através de um modelo linear generalizado misto, no qual o preditor linear contém parâmetros de efeitos fixos e de efeitos aleatórios.

Pode-se também usar a metodologia de equação de estimação generalizada, a qual permite incluir no modelo uma matriz de correlação de trabalho, a fim de obter estimadores consistentes mesmo que a matriz de correlação utilizada seja mal especificada. Essa abordagem é feita para a distribuição marginal ao invés da condicional dada as variáveis, embora a distribuição condicional possa ser mais apropriada para alguns casos.

O objetivo deste trabalho consiste em apresentar as metodologias estatísticas de modelo linear generalizado, modelo linear generalizado misto e equação de estimação generalizada para o tempo de germinação de sementes de mamona, levando em conta a proporção de sementes germinadas, visando contribuir com os estudos que têm sido realizados com sementes dessa matéria-prima utilizada na produção de biodiesel. Os dados de sementes de mamona da cultivar AL Guarany 2002 utilizados neste trabalho foram obtidos pela engenheira agrônoma Líbia Belisário da Silva (SILVA, 2007), entre novembro de 2005 e junho de 2006, na área de experimentação da Faculdade de Ciências Agrônomicas – UNESP, Campus de Botucatu.

## **4 REVISÃO BIBLIOGRÁFICA**

### **4.1 O Biodiesel**

No Brasil, a trajetória do biodiesel começou a ser delineada com as iniciativas de estudos pelo Instituto Nacional de Tecnologia, na década de 20 e ganhou destaque em 1980, com os trabalhos do professor Expedito José de Sá Parente (PARENTE, 2003), da Universidade Federal do Ceará, que patenteou o primeiro biodiesel brasileiro (PI – 8007957), e com a criação do Programa Nacional de Óleos Vegetais para Fins Energéticos (Pró-óleo), pela Resolução nº 7 do Conselho Nacional de Energia. O objetivo do programa Pró-óleo era promover a substituição de até 30% de óleo diesel apoiado na produção de soja, amendoim, colza e girassol. Devido à estabilização dos preços do petróleo e à entrada do Pró-álcool, juntamente com o custo elevado da produção e esmagamento das oleaginosas, o programa foi desacelerado (SEBRAE, 2007).

Com a guerra do Golfo Pérsico, em 1990, os EUA perderam dois grandes fornecedores de petróleo, o Iraque e o Kuwait, fazendo com que as cotações do barril de petróleo atingissem preços muito elevados. Em 2001, com os ataques terroristas em Nova Iorque e com a crise institucional na Venezuela, novamente o preço disparou. Esses acontecimentos juntamente com as pressões ambientais, fez com que muitos países, inclusive o Brasil, pensassem em alternativas como o biodiesel.



Acompanhando o movimento mundial de iniciativas favoráveis ao combate do efeito estufa, o Brasil introduziu, a partir de 2005, o biodiesel na matriz energética com as condições de mercado e a produção regulamentadas pela Lei nº 11.097/05. Essa Lei estabelece percentuais mínimos de mistura de biodiesel ao diesel, sendo de 2% no período autorizativo de 2005 a 2007, 2% no período obrigatório de 2008 a 2012, e 5% no período obrigatório de 2013 em diante (SEBRAE, 2007). Em janeiro de 2010, o governo brasileiro antecipou em 3 anos a obrigatoriedade da mistura de 5% de biodiesel ao diesel.

As estimativas de volumes previstos são de 1 bilhão de litros anuais de B2 (1 milhão de m<sup>3</sup>), na forma obrigatória nos anos de 2008 a 2012 e de 2,4 bilhões de litros anuais de B5 (2,4 milhões de m<sup>3</sup>), na forma de mistura de 5% de biodiesel e 95% de óleo diesel a partir de 2013 (SEBRAE, 2007).

A aplicação geral do biodiesel é como combustível, podendo ser utilizado em tratores na agricultura, caminhões no transporte de cargas, ônibus, trens, navios e também em geradores de energia. Suas principais vantagens são: exaustão menos ofensiva; não requer armazenamento especial; pelo fato de ter maior ponto de fusão, seu transporte é mais seguro; requer modificações mínimas para operar em motores já existentes; é renovável, contribuindo para a redução do dióxido de carbono; pode ser usado sozinho ou misturado em qualquer quantidade com o diesel de petróleo; aumenta a vida útil dos motores por ser mais lubrificante; é biodegradável e não tóxico.

O Brasil possui regiões distintas que possibilitam o cultivo de um número maior de culturas com propósito energético. Neste contexto, as características climáticas, culturais e tecnológicas de cada região condicionam o tipo de cultura a ser desenvolvida.

A mamona, o dendê, o babaçu, o girassol e outras oleaginosas das quais é extraído e fabricado o biodiesel, são cultivados no Brasil por pequenos e médios agricultores. Assim, a necessidade de um aumento na produção pode gerar também um aumento na renda dessas famílias e um aumento do emprego no campo (FREIRE et al., 2001).

A Petrobras Biocombustível concluiu, no dia 15 de novembro de 2009, todo o processo tecnológico que permite à empresa produzir biodiesel a partir do óleo de

mamona, dentro das especificações técnicas da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), que estipula o limite de 30% do óleo de mamona em cada litro de biodiesel. O biodiesel foi obtido com uma mistura de 30% de óleo de mamona e 70% de óleo de girassol, ambos produzidos pela agricultura familiar nos programas de suprimento de oleaginosas da empresa (PETROBRAS, 2009).

As pesquisas que levaram a Petrobras (Petróleo Brasileiro S. A.) ao domínio tecnológico do processo de produção do biodiesel a partir da mamona foram desenvolvidas pelo Centro de Pesquisa da Petrobras (Cenpes), na Usina de Guamaré, no Rio Grande do Norte, unidade experimental da companhia para o desenvolvimento do processo de produção do biodiesel.

## **4.2 Caracterização da mamoneira**

A mamoneira (*Ricinus communis L.*) é uma planta de origem tropical, possivelmente da Etiópia, leste de África. Resistente à seca e exigente em calor e luminosidade (20–30°C), requer para o seu crescimento e desenvolvimento pelo menos 500 mm de precipitação anual, equivalente a 5000 m<sup>3</sup> por hectare. Pode ser produzida em altitudes que variam de 300 a 1500 metros e semeada em vários tipos de solo, exceto nos solos muito argilosos, sujeitos a encharcamento, salinos e/ou sódicos, com elevado teor de sódio trocável.

É uma oleaginosa de alto valor, tanto econômico como social. No Brasil, é uma cultura produzida tradicionalmente em pequenas e médias propriedades, gerando emprego e renda em razão de suas inúmeras possibilidades de aplicação na área industrial, além da perspectiva de potencial energético na produção de biodiesel, tornando-se um agronegócio bastante promissor.

O óleo de mamona é o melhor óleo vegetal para fins industriais, pois não muda as suas características em altas, baixas ou em variações bruscas de temperatura. Do óleo de mamona são obtidos cosméticos, lubrificantes, plastificantes, detergentes, graxas, tintas, vernizes, poliuretanos, defensivos agrícolas, próteses humanas, entre outros.

Existem várias cultivares de mamoneira disponíveis para o plantio. Elas diferem em porte, deiscência dos frutos, tipos dos cachos e outras características. Nesse trabalho, utilizamos a cultivar AL Guarany 2002.

As plantas da cultivar AL Guarany 2002 têm altura de 1,60 a 2,60 m, caule de coloração roxo-avermelhada, frutos indeiscentes, sementes com coloração predominantemente marrom-escura com estrias cinza-claras. O peso médio de 1000 sementes é de 460 g, o teor de óleo da semente é de 47 a 48% e a produtividade média é de 1000 a 2000 kg de grãos por hectare, conforme a época do plantio e a tecnologia empregada. O ciclo médio é de 180 dias (AMARAL, 2004).

A utilização da mamona para a produção de biodiesel é uma realidade em nosso país, entretanto a escassez e a baixa qualidade das sementes utilizadas são entraves para a expansão da cultura, pois o cultivo ainda é realizado com sementes dos próprios agricultores e apresentam grande heterogeneidade.

Assim, pesquisas sobre a germinação de sementes de mamoneira são essenciais para que a cultura se estabeleça como uma boa alternativa agrícola para a produção de biodiesel frente a outras opções, como a soja, o amendoim e o girassol, que possuem uma tecnologia de produção no campo mais aprimorada.

A mamoneira também difere dessas espécies por não possibilitar a colheita mecânica na maioria das cultivares disponíveis no mercado e apresentar hábito de crescimento indeterminado, ou seja, a emissão da inflorescência e a fertilização são prolongadas, e ocorrem junto com o desenvolvimento da planta. Assim, a produção de sementes ocorre num período longo, tornando a operação de colheita mais dispendiosa e que demanda mão de obra para a cultura, por causa da necessidade de repetir o processo várias vezes durante um ciclo.

O conhecimento do processo de maturação é importante para o estabelecimento do ponto ideal de colheita, momento em que as sementes apresentam melhor qualidade fisiológica. O ponto de maturidade fisiológica depende da espécie, pois é afetado por fatores genéticos e, também, por fatores ambientais como a temperatura e a umidade relativa do ar. A maturação das sementes pode ser acelerada por temperaturas relativamente

altas, baixa umidade no solo e do ar. Fatores inversos a esses podem atrasar o processo de maturação (GEORGE, 1985). As sementes de mamona colhidas ainda imaturas apresentam qualidade física e fisiológica inferior, bem como massa, quantidade de óleo, germinação e vigor menores. Além disso, os frutos verdes constituem-se em meio propício para o desenvolvimento de fungos e precisam ser colocados ao sol para secar (MALLARD, 1961).

Em espécies que apresentam o florescimento contínuo como a mamoneira, a colheita dos frutos seguidos de um período de repouso pós-colheita pode trazer vantagens; esses frutos apresentam desuniformidade nos estágios de maturação e o repouso permitiria que os frutos imaturos completassem o processo. Dessa forma, um número menor de colheitas poderia ser praticado colhendo-se frutos em estágios de desenvolvimento distintos, mas próximos da maturidade fisiológica.

Na maior parte da área plantada com mamoneira no Estado de São Paulo são utilizadas cultivares indeiscentes, destacando-se a AL Guarany 2002, desenvolvida pelo Departamento de Sementes, Mudas e Matrizes da CATI, pois apresenta produtividade média de 1000 a 2500 kg/ha, ciclo de 180 dias, relativamente curto até a colheita do racemo terciário, porte médio de 1,60 a 2,60 m, o que facilita a colheita, resistência à seca e teor de óleo de 47 a 48% em suas sementes.

No mercado mundial, a Índia é o principal país produtor de mamona, seguido da China e do Brasil com, respectivamente, 51%, 35% e 8% da produção mundial do produto (FAO, 2009). O Brasil já foi o maior produtor mundial de mamona (573 mil toneladas em 1974) e o maior exportador do seu óleo (há algumas décadas). Em 1996, a produção nacional foi de 122 mil toneladas. No nordeste semiárido brasileiro concentra-se a produção nacional (80%), sendo de grande importância para a economia dessa região por ser resistente à seca, ser fixadora de mão de obra bem como geradora de emprego e de matéria-prima.

A mamoneira (Figura 1), cujo nome científico é *Rícinus communis* L., Classe *Dicotyledoneae*, Família *Euforbiaceae*, também é chamada popularmente de carrapateira, baforeira e бага. Segundo a cultivar, a mamoneira pode ter de 1,8 m até acima de 5 m de altura bem como cor da folha e caule, tamanho da semente e conteúdo de óleo variáveis. Em termos gerais na escolha da cultivar para plantio, além da adaptabilidade à

região, deve-se levar em consideração a produtividade, precocidade, deiscência do fruto, uniformidade de maturação, porte, entre outras características.



Figura 1: Plantação de mamona.

Possui raízes laterais e uma pivotante que vai a 1,5 m de profundidade, caule redondo, liso, esverdeado e coberto com cera, folhas verde-escuro, grandes, com 5 a 11 lóbulos, flores em racemo (cacho) terminal com flores masculinas (baixo), femininas e hermafroditas, com pólen viável por uma semana. Fruto e cápsula tricoca deiscente ou indeiscente, semente com cor e tamanho variados, com 40 a 49% de óleo que tem como componente maior o ácido ricinoleico.

Os restos culturais da mamoneira podem devolver ao solo 20 toneladas de biomassa; as folhas podem servir de alimento para o bicho-da-seda. A haste (caule) pode fornecer celulose para a fabricação de papel além de ser matéria-prima para a fabricação de tecidos grosseiros. Da semente extrai-se óleo de rícino tido como dos mais versáteis, que é produto renovável e barato tendo mais de 400 aplicações industriais. Resultante do

esmagamento da semente, a torta de mamona tem uso agrícola pela riqueza em nitrogênio (SEAGRI, 2009).

A mamoneira é planta anual (ciclo 250 dias) a semiperene (5 anos), de clima tropical e subtropical, precisa de chuvas regulares no início do período e de período seco na maturação dos frutos. Não suporta geada, ventos fortes frequentes e nebulosidade.

Requer temperatura entre 20 e 26°C, chuvas entre 600 e 700 mm anuais (mínimo de 400 mm), dias longos (com 12 horas de duração como mínimo) em altitude entre 300 e 1500 m. Em clima temperado, a planta desenvolve-se, mas tem a produção de óleo prejudicada. Na Bahia, planta-se mamona de abril a junho e de outubro a dezembro, segundo a zona agrícola (SEAGRI, 2009).

Devido às medidas governamentais que estimularam a cultura da mamona, surgiram várias pesquisas na área da produção de sementes para verificar a potencialidade da espécie e melhorar a qualidade do produto.

Carneiro e Pires (1983) verificaram a influência de temperaturas e substratos na capacidade germinativa de sementes de mamona da variedade Guarani. O experimento foi montado de acordo com um fatorial 4x3, com temperaturas de 20–35°C e de 20–30°C, alternadas, e de 30°C e 25°C, constante, em rolo de papel, rolo de pano e rolo de papel e pano. Através da análise de variância, seguida do teste de Tukey ao nível de 1% de significância, verificou-se nos três substratos que o poder germinativo das sementes foi o mesmo na temperatura de 30°C constante. Sementes que germinaram em temperaturas de 20–35°C alternadas tiveram maior poder germinativo quando foi usado o rolo de papel e pano.

Severino et al. (2004) utilizaram análise de variância para avaliar se a posição em que a semente de mamona é colocada possui efeito sobre o tempo necessário para a emergência e sobre o percentual de germinação. Concluíram que a posição de plantio da semente não influenciou o percentual de emergência nem o número de dias para emergência de 50% das plântulas de mamoneira da cultivar BRS 149 Nordestina plantadas em um solo arenoso e pouco estruturado.

Santos et al. (2006) verificaram a germinação de sementes de mamona das cultivares AL Guarany, Guarany, IAC 80, IAC 226 e Nordestina. Foi feita a análise de

variância para cada cultivar, sendo que para algumas cultivares, os tratamentos que tiveram porcentagem de germinação zero foram descartados da análise estatística. Para a comparação de médias das cultivares, que não apresentaram problemas de homogeneidade das variâncias, foi utilizado o teste de Tukey, com nível de significância de 5%. Os resultados obtidos permitiram concluir que a germinação de sementes de mamona deve ser realizada em germinadores do tipo Mangelsdorf, regulado à temperatura de 25°C.

Cartaxo et al. (2008) estudaram a germinação de sementes de mamona da cultivar BRS Energia. O delineamento experimental utilizado foi inteiramente ao acaso, no esquema fatorial 2x2x2, com quatro repetições. Os tratamentos foram representados por dois tipos de sementes (sem e com casca), sendo que metade destes materiais permaneceu sem tratamento químico e a outra metade foi tratada com fungicida. No teste de germinação das sementes de cada tratamento, foram utilizados dois substratos: papel Germitest e areia. Os resultados obtidos foram submetidos à análise de variância e as médias comparadas pelo teste de Tukey, ao nível de 5% de significância. Observou-se que no substrato de areia, a maior porcentagem de germinação (91,5%) ocorreu nas sementes com casca, superando o valor de 61% de germinação obtida pelas sementes sem casca. O desempenho germinativo no substrato rolo de papel foi mais eficiente para as semente sem casca e as sementes tratadas com fungicida apresentaram maior porcentagem de germinação.

Mendes et al. (2009) estudaram a eficiência de diferentes tratamentos pré-germinativos no desempenho de sementes de mamona, utilizando-se cinco lotes da cultivar Guarani. Os dados obtidos para cada lote e tratamento foram submetidos à análise de variância e as médias dos tratamentos foram comparadas pelo teste de Tukey. Em geral, todos os tratamentos pré-germinativos contribuíram para aumentar a porcentagem de germinação das sementes de mamona em relação à testemunha. Os tratamentos mais eficientes para acelerar a germinação das sementes foram a escarificação com lixa e a remoção da carúncula ou de todo o tegumento. Considerando a praticidade de aplicação, a escarificação com lixa pode ser recomendada para acelerar e aumentar a germinação das sementes de mamona.

Neste trabalho, os dados de germinação de sementes de mamona serão analisados usando as metodologias de Análise de Sobrevivência, Modelo Linear Generalizado, Modelo Linear Generalizado Misto e Equação de Estimção Generalizada.

### 4.3 Análise de Sobrevivência

Em alguns estudos, a variável resposta é geralmente, o tempo até a ocorrência de um evento de interesse. Exemplos incluem o tempo de vida de indivíduos acometidos de determinada doença, tempo de vida de espécies florestais ou o tempo até a germinação da semente. Esse intervalo de tempo até a ocorrência da germinação da semente é chamado de tempo de sobrevivência ou tempo de falha.

A principal característica de dados de sobrevivência é a presença de censura, que é a observação incompleta do tempo de germinação. Segundo Lawless (1982), existe censuras que são mais frequentes e que são encontradas tanto nos casos de censura à direita, onde o tempo de vida das observações é maior que o tempo em que ocorreu a censura, quanto de censura à esquerda, na qual o evento de interesse já ocorreu quando o indivíduo foi observado. Existe também a censura intervalar, muito comum quando se trabalha com intervalos de tempo (semanas, meses, anos). Nesse caso, sabe-se que o evento ocorreu no intervalo, mas não é possível determinar o tempo exato dessa ocorrência, gerando assim muitos empates nas respostas.

- Censura do tipo I: quando o estudo é encerrado após um período pré-estabelecido de tempo, de forma que as sementes que não germinarem até este instante serão censuradas. Na figura 2, as sementes C e E foram censuradas.

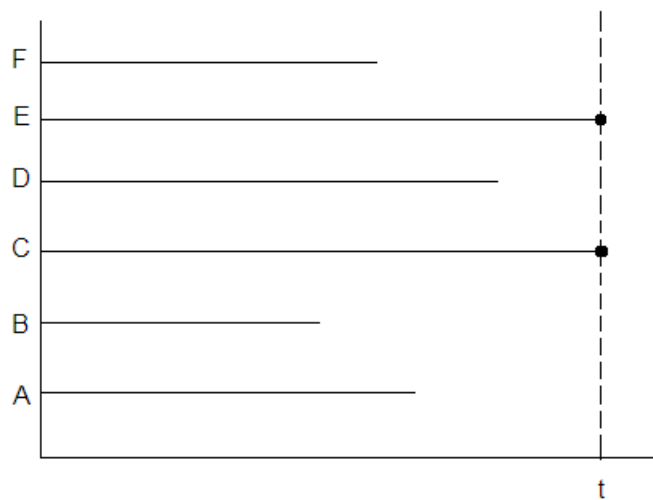


Figura 2: Censura do tipo I.



- Censura do tipo II: quando o estudo é encerrado após ter ocorrido o evento de interesse (germinação) em um número pré-estabelecido de sementes, de forma que as sementes que não germinarem até então serão censuradas. Na figura 3, supondo que o número pré-estabelecido de sementes germinadas seja 4, as sementes censuradas foram A e C.

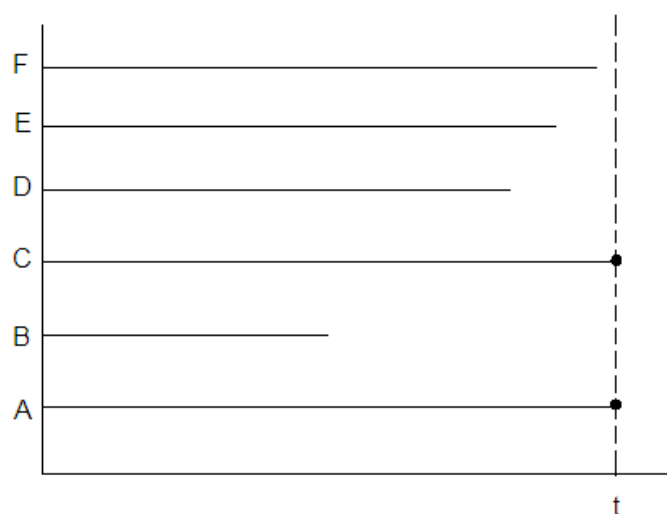


Figura 3: Censura do tipo II.

- Censura do tipo aleatório: quando a semente é retirada do estudo por causas alheias ao experimento, sem ter ocorrido a germinação. As sementes podem entrar no estudo em diferentes tempos. Na figura 4, as sementes B, C e F foram censuradas.

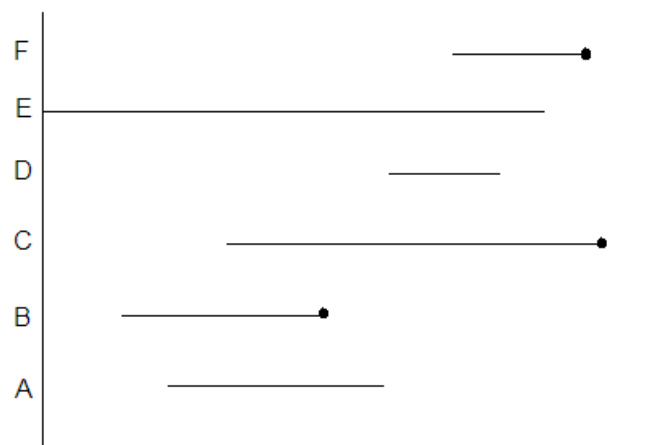


Figura 4: Censura do tipo aleatório.

No processo de coleta de dados de tempo de vida, pode não ser possível observar o tempo exato de ocorrência da germinação. É o que ocorre quando as observações são feitas em intervalos de tempo; ao final do intervalo, sabe-se se a germinação ou a censura ocorreu, mas não se pode determinar o momento exato de sua ocorrência. Neste caso, os dados são censurados à esquerda e à direita, constituindo a censura intervalar, e o tempo de germinação é tratado como uma variável aleatória discreta (LAWLESS, 1982).

- Censura intervalar: quando o estudo é encerrado após um período pré-estabelecido de tempo e as sementes que não germinarem em cada intervalo, serão censuradas. Na Figura 5, todas as sementes foram censuradas no primeiro intervalo; no segundo e no terceiro intervalo, temos 2 empates. A semente C foi censurada em todos os intervalos.

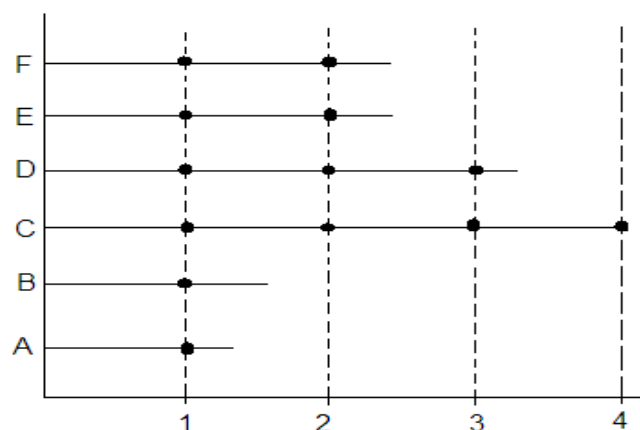


Figura 5: Censura intervalar.

Uma distribuição de tempos de sobrevivência pode ser caracterizada por uma função de sobrevivência, representada por  $S(t)$ , e é definida como a probabilidade de um indivíduo sobreviver por mais do que um determinado tempo  $t$ . Se  $T$  é uma variável aleatória contínua não negativa, que representa o tempo de sobrevivência, com função densidade de probabilidade  $f(x)$ ,  $x > 0$ , então:

$$S(t) = P(T \geq t) = \int_t^{\infty} f(x)dx, \quad (1)$$

com  $S(0) = 1$  e  $S(\infty) = 0$ .

A função de risco,  $h(t)$ , também chamada de taxa de falha, é definida como o risco instantâneo de um indivíduo sofrer o evento no intervalo de tempo  $(t, t + \Delta t)$ , dado que ele sobreviveu até o tempo  $t$ , e é expressa por:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (2)$$

A função de risco está relacionada com a função de sobrevivência através da expressão:

$$h(t) = \frac{f(t)}{S(t)}. \quad (3)$$

As funções de sobrevivência e de risco são inversamente proporcionais, isto é, quando o risco aumenta, a probabilidade de sobrevivência diminui e vice-versa.

Pode-se também relacionar a função de sobrevivência com a função de risco acumulado  $H(t)$ . A função de risco acumulado (ou taxa de falha acumulada) mede o risco de ocorrência do evento até um determinado tempo  $t$  e é dada pela expressão:

$$H(t) = \int_0^t h(x)dx = -\ln[S(t)]. \quad (4)$$

São diversos os métodos utilizados para a análise de dados de sobrevivência, dentre os quais podemos citar a tabela de vida, o estimador de Kaplan-Meier (KAPLAN; MEIER, 1958), o modelo exponencial, o modelo de Weibull (WEIBULL, 1951), os modelos de regressão de Cox (COX, 1972), os modelos para dados discretos, entre outros. Esses métodos podem ser classificados em duas grandes categorias: os métodos descritivos e os modelos de regressão.

Os métodos descritivos são utilizados quando o objetivo é simplesmente conhecer a função de sobrevivência de uma população. Esta função de sobrevivência pode ser estimada através de métodos paramétricos ou não paramétricos.

Em geral, os métodos paramétricos são mais eficientes que os não-paramétricos, no entanto, nem sempre é possível encontrar uma distribuição de probabilidade para a variável estudada. Esta dificuldade torna-se ainda maior quando se trata de dados de tempo de vida, os quais, devido à presença de censura, são dados incompletos. Em decorrência disso, os métodos não paramétricos, como a tabela de vida e o estimador de Kaplan-Meier, são bastante utilizados em análise de sobrevivência.

Os modelos de regressão consideram a relação entre o tempo de vida e outros fatores, nas situações que envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. O objetivo de um modelo de regressão é estimar o efeito das variáveis independentes (covariáveis) sobre o tempo de sobrevivência.

De acordo com a distribuição que acreditamos ter a variável resposta  $Y$ , e de acordo com a função escolhida para relacioná-la com as covariáveis, identificamos o modelo de regressão como Linear, Poisson, Logístico, Exponencial, etc.

#### **4.3.1 Estimação não-paramétrica**

O objetivo em análise de sobrevivência é estimar a função de sobrevivência  $S(t)$  da variável em estudo. Uma metodologia interessante para tratar esse problema é a estimação não-paramétrica, pois neste caso não se impõe um modelo teórico para o tempo de sobrevivência.

#### 4.3.1.1 Estimador de Kaplan-Meier

O método do produto-limite de se estimar a função de sobrevivência, também chamado método de Kaplan-Meier, proposto por Kaplan e Meier (1958), é uma técnica não-paramétrica bastante utilizada.

Sejam  $t_1 < t_2 < \dots < t_n$  os tempos onde ocorreram os eventos;  $R(t_i)$  o número de observações sob risco e  $N(t_i)$  o número observado de eventos no tempo  $t_i$ . A cada tempo  $t_i$  em que houver um evento, a probabilidade de sobrevivência será estimada pelo número de sobreviventes até aquele tempo ( $R(t_i) - N(t_i)$ ) sobre os que estavam em risco naquele tempo  $R(t_i)$ . O estimador da função  $S(t)$  é o produto das probabilidades de sobrevivência a cada tempo  $t_i \leq t$ :

$$\hat{S}(t) = \prod_{\substack{i=1 \\ t_i \leq t}}^n \frac{R(t_i) - N(t_i)}{R(t_i)}. \quad (5)$$

Para construir intervalos de confiança e testar hipóteses para  $S(t)$ , devemos obter a variância desse estimador. O estimador da variância mais utilizado para  $S(t)$  é o estimador de Greenwood (GREENWOOD, 1926), cuja expressão é:

$$\hat{Var}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{i: t_i \leq t} \frac{N(t_i)}{R(t_i)(R(t_i) - N(t_i))}. \quad (6)$$

Portanto, um intervalo aproximado de  $100(1 - \alpha)\%$  de confiança para  $S(t)$ , em um certo tempo  $t$ , é dado por:

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\hat{Var}(\hat{S}(t))}. \quad (7)$$

Entretanto, em valores extremos de  $t$ , este intervalo de confiança pode apresentar limite inferior negativo ou limite superior maior que 1, o que é incoerente com a definição de probabilidade. Isso pode ser resolvido utilizando uma transformação sugerida por Kalbfleisch e Prentice (1980) para  $S(t)$ , dada por  $\hat{U}(t) = \log(-\log(\hat{S}(t)))$ .

Assim, um intervalo aproximado de  $100(1 - \alpha)\%$  de confiança para  $S(t)$  é dado por:

$$\left[ \hat{S}(t) \right]^{\exp\left[\pm z_{\alpha/2} \sqrt{\hat{Var}(\hat{U}(t))}\right]}, \quad (8)$$

onde

$$\hat{Var}(\hat{U}(t)) = \frac{\sum_{i: t_i \leq t} \frac{N(t_i)}{R(t_i)[R(t_i) - N(t_i)]}}{\left[\log(\hat{S}(t))\right]^2}. \quad (9)$$

#### 4.3.2 Modelos para tempo de vida discreto

Considere que os tempos de vida,  $T_i$ ,  $i = 1, \dots, n$ , onde  $n$  é o tamanho da amostra, são agrupados em  $k$  intervalos,  $I_i = [a_{i-1}, a_i)$ ,  $i = 1, \dots, k$ , com  $0 = a_0 < a_1 < \dots < a_k = \infty$ , e assumamos que todas as censuras ocorrem no final de cada intervalo. Seja  $D_i$  o conjunto de sementes que germinaram no intervalo  $I_i$ ,  $R_i$  o conjunto de sementes sob risco no início de  $I_i$  e  $\Delta_{li}$  a variável indicadora de falha da  $l$ -ésima semente em  $I_i$ , a qual assume valor 0 quando a  $l$ -ésima observação é censurada no  $I_i$ -ésimo intervalo e assume valor 1 em caso contrário. Seja  $p_l(\mathbf{x}_l)$  a probabilidade da  $l$ -ésima semente germinar até  $a_i$  dado que não germinou até  $a_{i-1}$  e o vetor de variáveis regressoras  $\mathbf{x}_l$ . Assim a função de verossimilhança é dada por:

$$\prod_{i=1}^k \left( \prod_{l \in D_i} p_i(\mathbf{x}_l) \prod_{l \in R_i - D_i} [1 - p_i(\mathbf{x}_l)] \right) = \prod_{i=1}^k \prod_{l \in R_i} [p_i(\mathbf{x}_l)]^{\Delta_{li}} [1 - p_i(\mathbf{x}_l)]^{(1 - \Delta_{li})}. \quad (10)$$

A expressão (10) corresponde à função de verossimilhança de uma variável aleatória com uma distribuição de Bernoulli, cuja variável resposta é  $\Delta_{li}$  e a probabilidade de sucesso é  $p_i(\mathbf{x}_l)$ . A probabilidade  $p_i(\mathbf{x}_l)$  pode ser modelada considerando o modelo de riscos proporcionais de Cox ou considerando o modelo Logístico.

#### 4.3.2.1 Modelo de riscos proporcionais de Cox

Adotando o modelo de riscos proporcionais de Cox para o tempo de vida  $T$ , no qual  $\boldsymbol{\beta}$  é o vetor de parâmetros associado às covariáveis e  $h_o(t)$  é a função de risco basal, a função de risco para  $T$  dado  $\mathbf{x}_l$  é dada por:

$$h(t | \mathbf{x}_l) = h_o(t) \exp(\mathbf{x}_l' \boldsymbol{\beta})$$

e a função de sobrevivência, por:

$$S(t | \mathbf{x}_l) = \exp \left[ - \int_0^t h(u | \mathbf{x}_l) du \right] = [S_o(t)]^{\exp(\mathbf{x}_l' \boldsymbol{\beta})} \quad (11)$$

sendo  $S_o(t)$  a função de sobrevivência de base.

O componente não-paramétrico  $h_o(t)$  é chamado de função de risco basal devido ao fato de  $h(t) = h_o(t)$  quando  $\mathbf{x}_l = 0$ .

Este modelo é dito semiparamétrico por não assumir uma distribuição estatística para a função de risco basal. Assume apenas que as covariáveis agem multiplicativamente sobre o risco e esta é a parte paramétrica do modelo. Ele é denominado modelo de riscos proporcionais pelo fato da razão entre o risco de ocorrência do evento para duas observações diferentes ser constante no tempo (CARVALHO et al.,2005).

Modelando  $p_i(\mathbf{x}_l)$  pela expressão (11), tem-se:

$$1 - p_i(\mathbf{x}_l) = \frac{S_0(a_i|\mathbf{x}_l)}{S_0(a_{i-1}|\mathbf{x}_l)}$$

ou ainda,

$$p_i(\mathbf{x}_l) = 1 - \left[ \frac{S_0(a_i|\mathbf{x}_l)}{S_0(a_{i-1}|\mathbf{x}_l)} \right]^{\exp(\mathbf{x}_l'\boldsymbol{\beta})}. \quad (12)$$

Com a expressão (12), obtém-se a seguinte função de verossimilhança para dados de tabela de vida com variáveis regressoras:

$$\prod_{i=1}^k \prod_{l \in R_i} \left\{ 1 - \left[ \frac{S_0(a_i|\mathbf{x}_l)}{S_0(a_{i-1}|\mathbf{x}_l)} \right]^{\exp(\mathbf{x}_l'\boldsymbol{\beta})} \right\}^{\Delta_{li}} \left\{ \left[ \frac{S_0(a_i|\mathbf{x}_l)}{S_0(a_{i-1}|\mathbf{x}_l)} \right]^{\exp(\mathbf{x}_l'\boldsymbol{\beta})} \right\}^{(1-\Delta_{li})}.$$

A estimação dos parâmetros é feita pelo método da máxima verossimilhança, sendo que o logaritmo da função de verossimilhança pode ser escrito como:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\rho}) = \sum_{i=1}^k \sum_{l \in R_i} \left[ \Delta_{li} \ln(1 - \rho_i^{\exp(\mathbf{x}_l'\boldsymbol{\beta})}) + (1 - \Delta_{li}) \ln(\rho_i^{\exp(\mathbf{x}_l'\boldsymbol{\beta})}) \right]$$



onde  $\rho_i = S_0(a_i | \mathbf{x}_i) / S_0(a_{i-1} | \mathbf{x}_i)$ .

Prentice e Gloeckler (1978) sugeriram o uso da reparametrização  $\gamma_i = \ln(-\ln(\rho_i))$ , que torna os  $\gamma_i$  irrestritos, e a convergência do processo iterativo de estimação dos parâmetros mais rápida.

O logaritmo da função de verossimilhança reparametrizada fica:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^k \sum_{l \in R_i} [\Delta_{li} \ln(1 - \exp(-\exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta}))) - (1 - \Delta_{li}) \exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta})]$$

e as equações de máxima verossimilhança dadas por:

$$\frac{\partial \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \beta_j} = \sum_{i=1}^k \sum_{l \in R_i} \left[ \frac{\Delta_{li} x_{lj} \exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta})}{\exp(\exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta})) - 1} - (1 - \Delta_{li}) x_{lj} \exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta}) \right] = 0$$

$$\frac{\partial \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma})}{\partial \gamma_i} = \sum_{l \in R_i} \left[ \frac{\Delta_{li} \exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta})}{\exp(\exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta})) - 1} - (1 - \Delta_{li}) \exp(\gamma_i + \mathbf{x}_l' \boldsymbol{\beta}) \right] = 0$$

com  $j = 1, \dots, p$  covariáveis e  $i = 1, \dots, k$ , que podem ser resolvidas pelo método iterativo de Newton-Raphson, apresentado no Apêndice A.

#### 4.3.2.2 Modelo Logístico

Adotando o modelo logístico, tem-se que:

$$p_i(\mathbf{x}_i) = 1 - [1 + \varphi_i \exp(\mathbf{x}_i' \boldsymbol{\beta})]^{-1}, \quad (13)$$

em que  $\varphi_i = p_i(0)/(1 - p_i(0))$ , para  $i = 1, \dots, k$ .

Assim, a função de verossimilhança pode ser escrita como:

$$L(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \prod_{i=1}^k \prod_{l \in R_i} \left\{ \frac{\varphi_i \exp(\mathbf{x}'_l \boldsymbol{\beta})}{1 + \varphi_i \exp(\mathbf{x}'_l \boldsymbol{\beta})} \right\}^{\Delta_{li}} \left\{ \frac{1}{1 + \varphi_i \exp(\mathbf{x}'_l \boldsymbol{\beta})} \right\}^{(1-\Delta_{li})}.$$

Fazendo a reparametrização  $\alpha_i = \ln(\varphi_i)$ , o logaritmo da função de verossimilhança é dado por:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^k \sum_{l \in R_i} \Delta_{li} (\alpha_i + \mathbf{x}'_l \boldsymbol{\beta}) - \ln(1 + \exp(\alpha_i + \mathbf{x}'_l \boldsymbol{\beta}))$$

e as equações de máxima verossimilhança são dadas por:

$$\frac{\partial \ln L(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_j} = \sum_{i=1}^k \sum_{l \in R_i} x_{lj} \Delta_{li} - \frac{x_{lj} \exp(\alpha_i + \mathbf{x}'_l \boldsymbol{\beta})}{1 + \exp(\alpha_i + \mathbf{x}'_l \boldsymbol{\beta})} = 0$$

$$\frac{\partial \ln L(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \alpha_i} = \sum_{i=1}^k \sum_{l \in R_i} \Delta_{li} - \frac{\exp(\alpha_i + \mathbf{x}'_l \boldsymbol{\beta})}{1 + \exp(\alpha_i + \mathbf{x}'_l \boldsymbol{\beta})} = 0$$

onde  $j = 1, \dots, p$  e  $i = 1, \dots, k$ , que podem ser resolvidas usando o processo iterativo de Newton-Raphson, apresentado no Apêndice A.

O ajuste dos modelos torna-se mais simples quando são utilizados os conceitos de modelos lineares generalizados. Whitehead (1989) ajustou o modelo de riscos

proporcionais de Cox para dados agrupados, utilizando um modelo linear generalizado com função de ligação complemento log-log. No caso do ajuste de um modelo Logístico, utilizou-se o modelo linear generalizado com função de ligação logística (CHALITA, 1997). Detalhes do modelo linear generalizado são dados a seguir.

#### 4.4 Modelo Linear Generalizado

Dentre os métodos de modelagem estatística, os mais populares são os métodos de regressão linear, regressão logística, regressão de Poisson e alguns métodos que lidam com dados de sobrevivência. Todos esses métodos apoiam-se em gerar um modelo matemático que melhor descreva a relação entre um desfecho e uma ou mais variáveis explanatórias. Cada um desses modelos de regressão pode ser expresso como um tipo particular do modelo linear generalizado.

Como definido por Nelder e Wedderburn (1972), a estrutura de um modelo linear generalizado (MLG) é formada por três componentes: um componente aleatório representado por um conjunto de variáveis aleatórias independentes  $Y_1, \dots, Y_n$  provenientes de uma mesma distribuição que faz parte da família exponencial na forma canônica com médias  $\mu_1, \dots, \mu_n$ ; um componente sistemático composto por covariáveis ou variáveis independentes  $x_1, \dots, x_p$  tais que produzam um preditor linear  $\eta$  e uma função de ligação, que relaciona as duas componentes citadas acima. Uma importante característica dos MLG é a suposição de independência, ou pelo menos de não correlação, entre as observações. Contudo, existem extensões de MLGs para dados correlacionados (LIANG; ZEGGER, 1986).

i) componente aleatório – as variáveis respostas  $Y_1, \dots, Y_n$  são independentes com  $E(Y_i) = \mu_i$  e seguem uma distribuição que pertence à família exponencial na forma canônica, isto é, a função densidade pode ser escrita como:

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right\} \quad (14)$$

sendo  $a_i(\cdot)$ ,  $b(\cdot)$  e  $c(\cdot)$  funções conhecidas e específicas para cada distribuição;  $\phi > 0$  um parâmetro de dispersão ou escala, suposto conhecido, que depende de um único parâmetro canônico ou natural  $\theta_i$ .

Pode-se mostrar que a média e a variância de  $Y_i$  são dadas, respectivamente, por:

$$E(Y_i) = b'(\theta_i) = \mu_i \quad \text{e} \quad V(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i), \quad (15)$$

sendo  $V(\mu_i)$  a função de variância.

A Tabela 1 mostra os componentes da família exponencial para algumas distribuições conhecidas.

Tabela 1: Componentes da família exponencial para algumas distribuições conhecidas

Distribuição	$a(\phi)$	$\theta$	$b(\theta)$	$c(y; \phi)$
Normal $N(\mu; \sigma^2)$	$\sigma^2$	$\mu$	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left[ \ln(2\pi\sigma^2) + \frac{y^2}{\sigma^2} \right]$
Binomial $\text{Bin}(m; \pi)$	1	$\ln\left(\frac{\pi}{1-\pi}\right)$	$m \ln(1 + e^\theta)$	$\ln\left(\frac{m}{y}\right)$
Poisson $P(\mu)$	1	$\ln(\mu)$	$e^\theta$	$-\ln(y!)$
Binomial Negativa $\text{BN}(\mu; k)$	1	$\ln\left(\frac{\mu}{\mu + k}\right)$	$-k \ln(1 - e^\theta)$	$\ln\left[\frac{\Gamma(k + y)}{\Gamma(k)y!}\right]$
Gama $G(\mu; \nu)$	$\nu^{-1}$	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\nu \ln(\nu y) - \ln[\gamma \Gamma(\nu)]$
Normal Inversa $\text{IG}(\mu; \sigma^2)$	$\sigma^2$	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{\frac{1}{2}}$	$-\frac{1}{2} \left[ \ln(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y} \right]$

Fonte: Cordeiro e Demétrio (2007).

Como exemplo, seja  $Y$  uma variável aleatória com distribuição normal de média  $\mu$  e variância  $\sigma^2$ , denotada por  $Y \sim N(\mu; \sigma^2)$ . A função de densidade de probabilidade de  $Y$  é:

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

onde  $-\infty < \mu < +\infty$ ,  $-\infty < y < +\infty$  e  $\sigma^2 > 0$ .

A expressão acima pode ser escrita na forma canônica:

$$f(y; \mu, \sigma^2) = \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right]\right\}$$

que comparada com (14) leva a

$$\theta = \mu \quad ; \quad a(\phi) = \sigma^2 \quad ; \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2} \quad ; \quad c(y; \phi) = -\frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{y^2}{\sigma^2}\right].$$

Portanto, a distribuição normal com média  $\mu$  e variância  $\sigma^2$  pertence à família exponencial.

Como outro exemplo, seja  $Y$  uma variável aleatória binomial baseada em  $m$  repetições, em que a probabilidade de sucesso em uma repetição é  $\pi$ , denotada por  $Y \sim \text{Bin}(m; \pi)$ . Sua função de probabilidade é expressa por:

$$f(y; \pi) = \binom{m}{y} \pi^y (1 - \pi)^{m-y}$$

onde  $0 < \pi < 1$  e  $y = 0, 1, 2, \dots, m$ .

Sua função de probabilidade pode ser escrita na forma:

$$f(y; \pi) = \exp \left\{ \ln \binom{m}{y} + y \ln \left( \frac{\pi}{1 - \pi} \right) + m \ln(1 - \pi) \right\}$$

obtendo-se:

$$\theta = \ln \left( \frac{\pi}{1 - \pi} \right) \rightarrow \pi = \frac{e^\theta}{1 + e^\theta} ; \quad a(\phi) = 1 \quad ; \quad b(\theta) = m \ln(1 + e^\theta) \quad ; \quad c(y; \phi) = \ln \binom{m}{y}$$

e portanto, a distribuição binomial pertence à família exponencial apresentada na equação (14).

ii) componente sistemático – as variáveis explanatórias ou explicativas  $\mathbf{x}'_i = [x_{i1}, \dots, x_{ip}]$ ,  $i = 1, 2, \dots, n$  que dão origem a um vetor de preditores lineares  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ , sendo  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ ,  $p < n$ , um vetor de  $p$  parâmetros desconhecidos a serem estimados, e  $\mathbf{X}$  é a matriz do modelo de dimensão  $n \times p$ , cujas linhas são dadas por  $\mathbf{x}'_i$ .

iii) função de ligação – faz a ligação entre o componente aleatório e o componente sistemático por meio de uma função conhecida  $g(\cdot)$ , monótona e diferenciável, que liga a média  $\mu_i$  ao preditor linear, isto é,  $g(\mu_i) = \eta_i = \mathbf{x}'_i \boldsymbol{\beta}$ . Quando a função de ligação é escolhida tal que  $g(\mu_i) = \theta_i$ , ela é denominada função de ligação canônica.

No modelo normal linear, a média e o preditor linear são idênticos, dado que  $\eta$  e  $\mu$  podem assumir quaisquer valores na reta real  $(-\infty; +\infty)$ . Assim, temos que uma ligação do tipo identidade ( $\eta = \mu$ ) é adequada para modelar dados normais. Se  $Y$  tem distribuição Poisson, com  $\mu > 0$ , a função de ligação adequada é a logarítmica ( $\eta = \ln(\mu)$ ), pois esta tem o domínio positivo e o contradomínio na reta real. Para modelos que assumem a distribuição binomial para a variável resposta, onde  $0 < \pi < 1$ , o domínio da função de ligação deve, necessariamente, estar no intervalo  $(0; 1)$ , enquanto que seu contradomínio é o intervalo  $(-\infty; +\infty)$ . Três funções, que são as mais utilizadas, garantem esta condição para o modelo binomial. São elas:

i) Logit

$$\eta = \ln\left(\frac{\pi}{1-\pi}\right)$$

ii) Probit

$$\eta = \Phi^{-1}(\pi),$$

em que  $\Phi(\cdot)$  é a função de distribuição acumulada da normal padrão.

iii) Complemento log-log

$$\eta = \ln[-\ln(1-\pi)]$$

Uma vez que o modelo tenha sido formulado, o próximo passo é estimar o vetor de parâmetros  $\beta$ , pelo método da máxima verossimilhança, e avaliar as estimativas. Considerando uma amostra aleatória de  $n$  observações de uma distribuição pertencente à família exponencial (14), a função de verossimilhança é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{y}) = \prod_{i=1}^n f(y_i; \theta_i, \phi) = \exp \left\{ \sum_{i=1}^n \left[ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right] \right\}$$

cujo logaritmo é:

$$l(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{y}) = \ln L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{y}) = \sum_{i=1}^n \left[ \frac{1}{a_i(\phi)} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi) \right]. \quad (16)$$

Derivando-se a função (16) em relação a  $\beta_j$ , usando-se a regra da cadeia, obtém-se a função escore:

$$U_j = \frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^n \frac{1}{a_i(\phi)} (y_i - \mu_i) \left( \frac{\partial \theta_i}{\partial \eta_i} \right) x_{ij} \quad (17)$$

sendo que

$$\frac{\partial \theta_i}{\partial \eta_i} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{V(\mu_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right).$$

As equações (17) são não lineares e, em geral, não possuem soluções exatas. Soluções aproximadas podem ser encontradas pelo processo iterativo de Newton-Raphson ou pelo método escore de Fisher. Pelo método escore de Fisher, usando notação matricial, tem-se:

$$\boldsymbol{\beta}^{(m+1)} = \left( \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)} \quad (18)$$



em que  $\mathbf{X}$  é a matriz de especificação do modelo,  $\mathbf{W}$  é a matriz diagonal dos pesos, cujos elementos da diagonal,  $w_{ii}^{(m)}$ , são dados por:

$$w_{ii}^{(m)} = \frac{1}{V(y_i)} \left[ \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^{(m)} \right]^2, \quad i = 1, 2, \dots, n$$

e  $\mathbf{z}^{(m)}$  é o vetor da variável dependente ajustada no passo  $m$ , com seus elementos dados por:

$$z_i^{(m)} = \hat{\eta}_i^{(m)} + (y_i - \hat{\mu}_i^{(m)}) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^{(m)}, \quad i = 1, 2, \dots, n.$$

O método usual para iniciar o algoritmo é especificar uma estimativa inicial  $\boldsymbol{\beta}^{(0)}$  e usando (18), iterá-la sucessivamente até obter convergência. Para efeito de passo inicial, pode-se especificar o vetor  $\boldsymbol{\beta}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\eta}}$ , sendo que  $\hat{\eta}_i = g(\hat{\mu}_i) = g(y_i)$ .

#### 4.4.1 Quase-verossimilhança

No método da máxima verossimilhança, a distribuição probabilística da qual os dados são provenientes, deve ser conhecida. Quando isso não ocorre, mas conhecemos a relação entre a média e a variância, podemos usar essa informação no processo de estimação. Wedderburn (1974) propôs o método da quase-verossimilhança, que requer apenas especificação de uma função de variância para a variável resposta e a relação funcional entre a resposta média e os parâmetros  $\boldsymbol{\beta}$ , e relaxa a suposição de que a distribuição, necessariamente, pertença à família exponencial.

Sejam  $n$  observações para a variável aleatória  $Y_i$ , tal que  $E(Y_i) = \mu_i$  e  $\text{Var}(Y_i) = \phi V(\mu_i)$ . Suponha também que  $\mu_i$  é uma função do conjunto de parâmetros de interesse  $\beta_1, \beta_2, \dots, \beta_p$ . A obtenção das estimativas dos parâmetros em  $\beta$  consiste na otimização do logaritmo da função de quase-verossimilhança  $Q$ , definida como:

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$

e, portanto,

$$\frac{\partial Q(\mu_i, y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)}.$$

#### 4.4.2 Qualidade do ajuste e seleção de modelos

A função desvio (*deviance*), proposta por Nelder e Wedderburn (1972), avalia a qualidade do ajuste de um modelo linear generalizado, ou seja, mede a discrepância entre o valor observado na amostra e aquele predito pelo modelo estimado. Sua expressão é dada por:

$$D(\hat{\mu}_i, y_i) = -2 \sum_{i=1}^n \{l(\hat{\mu}_i, y_i) - l(y_i, y_i)\} \quad (19)$$

sendo  $l(\hat{\mu}_i, y_i)$  o valor do logaritmo da função de verossimilhança do modelo sob pesquisa e  $l(y_i, y_i)$  o valor do logaritmo da função de verossimilhança do modelo saturado, que é o modelo para o qual os valores ajustados  $\hat{\mu}_i$  são iguais às respostas observadas  $y_i$ , ou seja, o modelo mais completo possível. Um valor pequeno para a função desvio indica que para um

número menor de parâmetros, obtém-se um ajuste tão bom quanto o ajuste no modelo saturado, que é o que contém mais parâmetros.

Lindsey (1997) demonstrou que  $D(\hat{\mu}_i, y_i)$  tem, assintoticamente, distribuição  $\chi^2$  com  $n-p$  graus de liberdade, sendo  $n$  o número de observações e  $p$  o número de parâmetros do modelo.

Na Tabela 2, têm-se as funções desvio (*deviances*) para algumas distribuições conhecidas.

Tabela 2: *Deviances* para algumas distribuições conhecidas

Distribuição	<i>Deviance</i>
Normal	$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2$
Binomial	$2 \sum_{i=1}^n \left[ y_i \ln \left( \frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \ln \left( \frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right]$
Poisson	$2 \sum_{i=1}^n \left[ y_i \ln \left( \frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]$
Binomial Negativa	$2 \sum_{i=1}^n \left[ y_i \ln \left( \frac{y_i}{\hat{\mu}_i} \right) + (y_i + k) \ln \left( \frac{\hat{\mu}_i + k}{y_i + k} \right) \right]$
Gama	$2 \sum_{i=1}^n \left[ \ln \left( \frac{\hat{\mu}_i}{y_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$
Normal Inversa	$\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$

Fonte: Cordeiro e Demétrio (2007).

Outra medida de discrepância do ajuste de um modelo a um conjunto de dados é a estatística Qui-Quadrado generalizada de Pearson, dada pela expressão:

$$Q_p = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \quad (20)$$

sendo  $V(\hat{\mu}_i)$  a função de variância estimada sob o modelo que está sendo ajustado aos dados. Para respostas com distribuição Normal,  $Q_p \sim \sigma^2 \chi_{n-p}^2$ , sendo esse resultado exato somente se  $\sigma^2$  for conhecido e a função de ligação for a identidade. Para as demais distribuições têm-se apenas resultados assintóticos.

O critério da informação de Akaike (AIC) proposto por Akaike (1974) pode ser usado na seleção de modelos e sua expressão é dada por:

$$AIC = -2l(\hat{\mu}_i, y_i) + 2p$$

sendo  $p$  o número de parâmetros. O critério consiste em encontrar o modelo tal que a quantidade acima seja minimizada.

#### 4.4.3 Análise de resíduos e diagnósticos

As técnicas usadas para análise de resíduos e diagnósticos nos MLGs são semelhantes às dos modelos lineares clássicos, com algumas alterações. Os tipos de resíduos mais utilizados para análise e diagnósticos dos MLGs são:

i) Resíduos ordinários

$$r_i = y_i - \hat{\mu}_i \quad (21)$$

ii) Resíduos de Pearson generalizados

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad (22)$$

iii) Resíduos de Pearson generalizados estudentizados internamente

$$r_i^{p'} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - \hat{h}_{ii})}} \quad (23)$$

onde  $h_{ii}$  é o  $i$ -ésimo elemento da diagonal da matriz de projeção  $\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X}(\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{1/2}$ .

iv) Componentes do desvio

$$r_i^D = \text{sinal}(y_i - \hat{\mu}_i) \sqrt{2[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]} \quad (24)$$

sendo  $\hat{\theta} = \hat{\theta}(\hat{\mu})$  e  $\tilde{\theta} = \tilde{\theta}(y)$  as estimativas do parâmetro canônico para os modelos saturado e corrente, respectivamente.

v) Componentes do desvio estudentizado internamente

$$r_i^{D'} = \frac{r_i^D}{\sqrt{1 - \hat{h}_{ii}}} \quad (25)$$

que é o resíduo dado em (24) padronizado, segundo McCullagh (1987), que usa as aproximações propostas por Cox e Snell (1968).

A análise de diagnóstico é uma etapa importante do ajuste de um modelo de regressão, pois nos auxilia na verificação de possíveis afastamentos das suposições feitas para o modelo e nos permite detectar observações extremas que interferem nos

resultados do ajuste. Entre as principais técnicas exploratórias para diagnósticos dos MLGs, podemos citar o estudo da diagonal principal da matriz de projeção  $H$  e a distância de Cook, úteis na identificação de pontos influentes, que são observações cuja omissão no conjunto de dados leva a mudanças significativas no ajuste do modelo.

Outra alternativa para se verificar se um modelo está bem ajustado aos dados é utilizar a técnica exploratória do gráfico de probabilidade normal (*q-q plot*). Esse gráfico é obtido plotando-se os valores de algum tipo de resíduo contra os percentis de uma distribuição Normal. Um modelo adequado indica que os pontos se distribuem em torno de uma reta.

#### **4.5 O Método da Equação de Estimação Generalizada**

Os modelos lineares generalizados são propostos para a análise de dados com apenas uma resposta por observação, os quais assumem independência entre as observações. Muitos autores têm proposto extensões de modelos lineares generalizados para a análise de dados correlacionados, por exemplo, Liang e Zeger (1986), Zeger e Liang (1986), Stram, Wei e Ware (1988), Moulton e Zeger (1989) e Zhao e Prentice (1990).

Para análise de dados com medidas repetidas, existe uma grande variedade de técnicas quando a variável resposta segue uma distribuição Normal. Entre elas, podemos destacar a análise de variância com medidas repetidas (NETER et al., 1996), análise de perfil, análise de curvas de crescimento (SINGER; ANDRADE, 1986) e modelos normais com efeitos aleatórios (LAIRD; WARE, 1982). O fato de existir uma versão multivariada da distribuição Normal com boas propriedades do ponto de vista inferencial facilita o estudo desses modelos. Porém, abandonando a suposição de normalidade da variável resposta, uma série de dificuldades pode surgir devido à escassez de distribuições multivariadas alternativas com tais propriedades (VENEZUELA, 2003).

Justamente por esse motivo, não são muitas as técnicas existentes na literatura para se trabalhar com distribuições não normais levando em consideração a dependência entre observações da mesma unidade experimental. Modelos de regressão com

respostas binárias são apresentados, por exemplo, por Kim et al. (2006) que descrevem uma aplicação de resposta binária em biometria, por Anderson e Aitkin (1985), Ochi e Prentice (1984) e por Verbeke e De Clercq (2006).

Prentice e Zhao (1991) utilizaram equações de estimação para obter estimativas consistentes dos parâmetros de regressão e de correlação e, neste caso, é necessário que tanto o modelo de regressão como a estrutura de correlação estejam corretamente especificados. Métodos para análise de dados que ignoram essa dependência entre as observações tendem a subestimar os erros-padrão (ZEGGER; KARIM, 1991; DIGGLE; LIANG; ZEGGER, 1994).

Liang e Zeger (1986) propuseram analisar dados com medidas repetidas utilizando modelos lineares generalizados. Eles obtêm estimativas consistentes para os parâmetros de regressão, mesmo que a estrutura de correlação esteja definida incorretamente, utilizando equações de estimação generalizadas (EEG), descritas nesta seção, e tratam os parâmetros de correlação como parâmetros de perturbação. As principais razões para considerar a abordagem EEG são que ela pode ser usada para análise de dados longitudinais, sejam eles discretos ou contínuos, sua teoria mostra propriedades importantes de consistência dos estimadores e que a metodologia está implementada no SAS (SAS INSTITUTE, 2008).

As EEGs são uma extensão multivariada da função de quase verossimilhança, inicialmente apresentada por Wedderburn (1974) e, mais tarde, por McCullagh e Nelder (1989). Essa função não exige conhecimento da distribuição paramétrica da variável resposta; é necessário especificar apenas a relação entre a média e a variância das observações, supondo alguma estrutura de correlação para os dados.

A metodologia EEG modela uma função conhecida da esperança marginal de uma variável dependente como uma função linear de uma ou mais variáveis explanatórias. Como na teoria de modelo linear generalizado, a variância é assumida ser uma função conhecida da média. Além disso, uma matriz de correlação para as observações de cada indivíduo é requerida. Essa metodologia resulta em estimadores consistentes dos coeficientes de regressão e de suas variâncias até mesmo sob fracas suposições acerca da atual

correlação entre as respostas numa mesma observação. Essa abordagem assume que apenas a resposta média da população, modelada como uma função das covariáveis, é o foco de interesse e evita a necessidade de distribuições multivariadas, assumindo apenas uma forma funcional para a distribuição marginal em cada ponto.

Seja  $y_{ij}$  a resposta do indivíduo  $i$  no tempo  $j$ , para  $i = 1, \dots, n$  e  $j = 1, \dots, t_i$ . Seja também  $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})'$  denotando um vetor  $p \times 1$  de variáveis explanatórias (covariáveis) associadas à  $y_{ij}$ . Se todas as covariáveis são constantes no tempo, então  $x_{i1} = x_{i2} = \dots = x_{it}$ . Note-se que  $y_{ij}$  e  $x_{ij}$  são valores perdidos quando as observações não são obtidas no tempo  $j$ .

O primeiro passo da abordagem EEG é relacionar a resposta marginal  $\mu_{ij} = E(y_{ij})$  a uma combinação linear das covariáveis :  $g(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}$ , onde  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  é um vetor  $p \times 1$  de parâmetros desconhecidos, que caracteriza o quanto a distribuição da resposta depende das variáveis explanatórias e  $g$  é uma função de ligação conhecida. Geralmente, as funções de ligação são a função logit,  $g(\mu_i) = \log(\pi/(1-\pi))$  para respostas binárias e a função logarítmica,  $g(\mu_i) = \log(\mu)$  para contagens Poisson.

O segundo passo é escolher a forma de uma matriz de correlação de trabalho  $\mathbf{R}(\boldsymbol{\alpha})$  de dimensão  $t_i \times t_i$  para cada  $\mathbf{y}_i = (y_{i1}, \dots, y_{it})'$ . O elemento  $(j, j')$  de  $\mathbf{R}_i(\boldsymbol{\alpha})$  é a correlação conhecida, hipotética ou estimada entre  $y_{ij}$  e  $y_{ij'}$ . Essa matriz de correlação pode depender de parâmetros desconhecidos  $\boldsymbol{\alpha}$ , o qual é o mesmo para todos os indivíduos. Assumindo-se que  $\mathbf{R}(\boldsymbol{\alpha})$  é conhecido, exceto para um número fixo de parâmetros  $\boldsymbol{\alpha}$  que devem ser estimados dos dados. Embora essa matriz de correlação possa diferir de indivíduo para indivíduo, geralmente usamos uma matriz de correlação  $\mathbf{R}(\boldsymbol{\alpha})$  que aproxima a dependência média entre observações repetidas sobre os indivíduos.

$\mathbf{R}$  é chamada de matriz correlação de trabalho porque com respostas não normais, a correlação atual entre as respostas do indivíduo pode depender dos valores médios. O método EEG produz estimativas consistentes dos coeficientes da regressão e suas variâncias, até mesmo com várias especificações da estrutura da matriz de covariância. Em



adição, a perda de eficiência de uma escolha incorreta de  $\mathbf{R}$  geralmente não tem consequência quando o número de indivíduos é grande.

Muitas possibilidades para a estrutura de correlação de trabalho têm sido sugeridas (LIANG; ZEGGER, 1986). As estruturas mais comuns de correlação de trabalho apresentadas pelos autores são:

#### i. Independente

Assume-se a matriz identidade,  $\mathbf{R}(\alpha) = \mathbf{I}_t$ , como matriz de correlação de trabalho. Assim, temos a suposição de que observações repetidas para um indivíduo são independentes. Nesse caso, resolver o EEG é o mesmo que ajustar o modelo de regressão usual para dados independentes.

#### ii. Permutável ou simétrica

A matriz de correlação de trabalho permutável assume que a correlação é constante entre quaisquer duas observações, isto é,  $R_{jj'} = 1$ , para  $j = j'$ ;  $R_{jj'} = \alpha$ , para  $j \neq j'$ . Essa é a estrutura de correlação assumida em um modelo de efeitos aleatórios com um intercepto aleatório para cada indivíduo (LAIRD; WARE, 1982). Exemplificando, para  $t_i = 3$ , temos:

$$\mathbf{R}(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix}$$

A estimativa para  $\alpha$  é dada por:

$$\hat{\alpha} = \hat{\phi} \sum_{i=1}^n \sum_{j>j'} \frac{\hat{r}_{ij} \hat{r}_{ij'}}{\sum_{i=1}^n \frac{1}{2} t_i (t_i - 1) - p}$$

sendo

$$\hat{\phi}^{-1} = \frac{\sum_{i=1}^n \hat{r}_{ij}^2}{\sum_{i=1}^n t_i - p}$$

e  $\hat{r}_{ij}$  é o resíduo de Pearson, dado por  $\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{V(\hat{\mu}_{ij})}}$ .

### iii. Não-estruturada

Quando a matriz de correlação é não-estruturada, isto é,  $R_{jj'} = 1$ , para  $j = j'$ ;  $R_{jj'} = \alpha_{jj'}$ , para  $j \neq j'$ , existem  $t_i (t_i - 1)/2$  parâmetros a serem estimados. Em particular, para  $t_i = 3$ , tem-se:

$$\mathbf{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & 1 & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & 1 \end{bmatrix}$$

havendo, portanto, 3 parâmetros a serem estimados. Esta estrutura é interessante quando não há muitos instantes de observação.

## iv. M-dependente

Com essa estrutura, a correlação depende das distâncias entre as medidas repetidas, podendo tender a zero para os espaçamentos maiores do que  $t$ . Neste caso, há  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_{t-1})'$  parâmetros a serem estimados, para os tempos  $j = 1, 2, \dots, t-1$ , e

$$\mathbf{R}(\alpha) = \begin{cases} 1, & \text{se } j = 0 \\ \alpha_j, & \text{se } j = 1, 2, \dots, t-1 \\ 0, & \text{se } j > t \end{cases}$$

Um estimador natural para  $\alpha_j$  é:

$$\hat{\alpha}_j = \hat{\phi} \sum_{i=1}^n \frac{\hat{r}_{ij} \hat{r}_{i,j+1}}{n-p}.$$

## v. Auto-Regressiva (AR-1)

Com uma estrutura de correlação auto-regressiva, as correlações também dependem das distâncias entre as medidas repetidas, sendo que as observações mais próximas tendem a ser mais correlacionadas, ou seja:

$$\mathbf{R}(\alpha) = \begin{cases} 1, & \text{se } j = j' \\ \alpha^{|j-j'|}, & \text{se } j \neq j' \end{cases}.$$

Para  $t_i = 3$ ,  $\mathbf{R}(\alpha)$  pode ser expressa por:

$$\mathbf{R}(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix}$$

A escolha da estrutura de correlação mais adequada em cada situação é a critério do pesquisador. Também em determinadas situações não existem grandes diferenças entre as estimativas obtidas segundo as diferentes estruturas de correlação. Horton e Lipsitz (1999) afirmaram que comparações de estimativas e erros-padrão a partir de várias estruturas de correlação podem indicar sensibilidade na não especificação correta da estrutura de variância.

O terceiro passo da abordagem EEG é estimar o vetor de parâmetros  $\beta$  e sua matriz de covariância. Primeiro, seja  $A_i$  a matriz diagonal de dimensão  $t \times t$  com  $V(\mu_{ij})$  como o  $j$ -ésimo elemento da diagonal. A matriz de variância-covariância para  $Y_i$  é  $V_i(\alpha) = \phi^{-1} A_i^{1/2} R_i(\alpha) A_i^{1/2}$ . A estimativa EEG de  $\beta$  é a solução da equação de estimação

$$U(\beta) = \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \beta} \right)' [V(\hat{\alpha})]^{-1} (y_i - \mu_i) = 0_p \quad (26)$$

onde  $\mu_i = (\mu_{i1}, \dots, \mu_{it})'$ ,  $0_p$  é o vetor de dimensão  $p \times 1$  com todos os elementos nulos e  $\hat{\alpha}$  é um estimador consistente de  $\alpha$ .

Essa equação de estimação é resolvida pela combinação do método da quase-verossimilhança para estimação de  $\beta$  com um método robusto para estimação de  $\alpha$  como uma função de  $\beta$ , como segue:

1. Dadas as atuais estimativas de  $\mathbf{R}(\alpha)$  e  $\phi$ , calculamos uma estimativa atualizada de  $\beta$ , usando iterativamente mínimos quadrados ponderados.

2. Dada a estimativa  $\hat{\boldsymbol{\beta}}$  de  $\boldsymbol{\beta}$ , calculamos os resíduos de Pearson padronizados

$$\mathbf{g}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{[V_i(\hat{\boldsymbol{\alpha}})]_{ij}}}.$$

3. Use os resíduos  $\mathbf{g}_{ij}$  para estimar consistentemente  $\boldsymbol{\alpha}$  e  $\phi$ .  
 4. Repita os passos 1 a 3 até obter convergência.

A variância de  $\hat{\boldsymbol{\beta}}$  é então estimada por  $\hat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \mathbf{M}_0^{-1} \mathbf{M}_1 \mathbf{M}_0^{-1}$ , onde

$$\mathbf{M}_0 = \sum_{i=1}^n \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' [\mathbf{V}(\hat{\boldsymbol{\alpha}})]^{-1} \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)$$

e

$$\mathbf{M}_1 = \sum_{i=1}^n \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' [\mathbf{V}(\hat{\boldsymbol{\alpha}})]^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) (\mathbf{y}_i - \boldsymbol{\mu}_i)' [\mathbf{V}(\hat{\boldsymbol{\alpha}})]^{-1} \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right).$$

Esse é um estimador robusto, que será sempre consistente mesmo que a matriz de correlação de trabalho seja definida incorretamente.

Testes de hipóteses para  $\boldsymbol{\beta}$  ou para subconjuntos de  $\boldsymbol{\beta}$  podem ser desenvolvidos através de estatísticas tipo Wald com a matriz de variância-covariância estimada.

Pan (2001) apresentou medidas para escolher a melhor matriz de correlação de trabalho e para selecionar covariáveis, baseadas no critério de informação de Akaike (AIC). Preisser e Qaqish (1996) apresentaram medidas para detectar observações influentes em MLGs com medidas repetidas.

Considerando-se que a suposição de normalidade dos resíduos de Pearson não necessariamente ocorre para respostas discretas, tais como respostas Poisson e binárias, Park et al. (1998) estudam os resíduos de Ascombe e *deviances* para a estimação dos parâmetros de correlação e comparam com os resíduos de Pearson. Os resultados mostram que a escolha do resíduo tinha pouco efeito (ou nenhum) nas propriedades das estimativas resultantes e, portanto, recomendam o uso dos resíduos de Pearson (COSTA, 2003).

Littell et al. (2006) recomendaram a estatística Qui-Quadrado Generalizada, dada pela expressão  $Q_g = \hat{\mathbf{r}}' [\mathbf{V}(\boldsymbol{\alpha})]^{-1} \hat{\mathbf{r}}$  como critério de verificação do ajuste do modelo. Se a divisão dessa estatística pelo grau de liberdade do resíduo for próximo a 1, tem-se um bom ajuste do modelo.

#### 4.6 Modelo Linear Generalizado Misto

Os Modelos Mistos são ferramentas poderosas na análise de dados agrupados, cujas observações são correlacionadas. Tais correlações podem estar associadas às parcelas ou às subparcelas dependendo do plano experimental adotado. O nome modelo misto vem do fato de que o modelo contém parâmetros de efeitos fixos,  $\boldsymbol{\beta}$ , e parâmetros de efeitos aleatórios,  $\mathbf{d}_i$ . A grande procura por modelos mistos pode ser explicada em função da flexibilidade oferecida pelos mesmos na modelagem das estruturas de correlação “intra-grupo”, que são de fundamental importância na análise de dados agrupados. Assim, é muito comum utilizar esses modelos em diversas áreas do conhecimento científico, tais como Agronomia, Biologia, etc.

A inclusão de um efeito aleatório para cada observação torna as estimativas dos efeitos das covariáveis mais consistentes, ou seja, ainda que estejam faltando algumas covariáveis importantes, as quais limitam a capacidade explicativa do modelo, pelo menos para as variáveis medidas pode-se estimar o parâmetro e erro padrão associado de forma mais precisa. Essa melhoria no ajuste do modelo não aumenta a nossa compreensão dos

mecanismos que influenciam o modelo, mas aumenta a confiança nas estimativas sobre o que foi possível medir (CARVALHO et al., 2005).

O modelo linear generalizado misto (MLGM) pertence à família de modelos para elemento-específico, que são modelos cujas medidas da variável resposta são assumidas independentes, dado um conjunto de parâmetros específicos para cada elemento na amostra. O MLGM assume que a correlação entre as respostas para um indivíduo seja resultante de uma heterogeneidade natural nos coeficientes de regressão entre os indivíduos e, a distribuição condicional das respostas dado um coeficiente individual, são variáveis aleatórias condicionalmente independentes, seguindo um modelo linear generalizado (DIGGLE et al., 2002; MOLENBERGHS; VERBEKE, 2005). Assim, tem-se:

$$\mathbf{Y}_i | \mathbf{d}_i \sim f_{\mathbf{Y}_i | \mathbf{d}_i}(\mathbf{y}_i | \mathbf{d}_i),$$

$$f_{\mathbf{Y}_i | \mathbf{d}_i}(\mathbf{y}_i | \mathbf{d}_i) = \exp \{ \phi^{-1} [\mathbf{y}_i \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)] + c(\mathbf{y}_i; \phi) \} \quad (27)$$

De (15) tem-se que  $\boldsymbol{\mu}_i = b'(\boldsymbol{\theta}_i)$ , portanto:

$$E(\mathbf{Y}_i | \mathbf{d}_i) = \boldsymbol{\mu}_i$$

é a média condicional da variável aleatória  $\mathbf{Y}_i$ , que é funcionalmente ligada à parte sistemática por uma função de ligação:

$$g[E(\mathbf{Y}_i | \mathbf{d}_i)] = g(\boldsymbol{\mu}_i) = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \mathbf{d}_i, \quad (28)$$

sendo  $\mathbf{x}'_i$  a  $i$ -ésima linha da matriz de delineamento associada aos efeitos fixos,  $\boldsymbol{\beta}$  é o vetor de parâmetros dos efeitos fixos,  $\mathbf{z}'_i$  é a  $i$ -ésima linha da matriz do modelo associada aos efeitos aleatórios e  $\mathbf{d}_i$  é o vetor dos efeitos aleatórios.

Assume-se uma distribuição aos efeitos aleatórios,  $\mathbf{d}_i \sim f_{d_i}(\mathbf{d}_i)$ , sendo usual atribuir a distribuição normal multivariada aos efeitos aleatórios, isto é,  $\mathbf{d}_i \sim N_q(\mathbf{0}; \mathbf{G})$ . Uma discussão mais aprofundada sobre os efeitos aleatórios não normais é dada por Lee e Nelder (1996), que usam distribuições conjugadas.

Considerando o modelo condicional (27), tem-se:

$$E(Y_i) = E[E(Y_i | \mathbf{d}_i)] = E[\mu_i] = E[g^{-1}(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{d}_i)]$$

que, em geral, não pode ser simplificado devido à presença de funções não lineares em  $g^{-1}(\cdot)$ .

A variância marginal de  $Y_i$  é dada por:

$$V(Y_i) = V[g^{-1}(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{d}_i)] + E\{a_i(\phi) V[g^{-1}(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{d}_i)]\}.$$

O uso de efeitos aleatórios também introduz uma correlação entre observações que tenham algum efeito em comum. Assumindo-se independência condicional dos elementos de  $\mathbf{Y}$ , tem-se:

$$\text{Cov}(\mathbf{Y}_i, \mathbf{Y}_t) = \text{Cov}[g^{-1}(\mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{d}_i), g^{-1}(\mathbf{x}'_t \boldsymbol{\beta} + \mathbf{z}'_t \mathbf{d}_t)].$$

Nesta análise deseja-se estimar os coeficientes dos efeitos fixos,  $\boldsymbol{\beta}$ , os parâmetros de  $f_{d_i}$  e, possivelmente, o parâmetro de dispersão  $\phi$ .



#### 4.6.1 Estimação dos parâmetros por máxima verossimilhança

A estimação de  $\beta$  e dos parâmetros de  $\mathbf{G}$  é feita através da maximização da função de verossimilhança marginal, obtida através da integração da função de verossimilhança sobre os efeitos aleatórios  $\mathbf{d}_i$ . A contribuição do  $i$ -ésimo elemento da função de verossimilhança é:

$$f(\mathbf{y}_i | \beta, \mathbf{G}, \phi) = \int \dots \int \prod_{t=1}^{n_i} f(y_{it} | \mathbf{d}_i, \beta, \phi) f(\mathbf{d}_i | \mathbf{G}) d(\mathbf{d}_i) \quad (29)$$

e a função de verossimilhança marginal para a amostra contendo os  $N$  elementos é dada por:

$$L(\beta, \mathbf{G}, \phi) = \prod_{i=1}^N f(\mathbf{y}_i | \beta, \mathbf{G}, \phi) = \prod_{i=1}^N \int \dots \int \prod_{t=1}^{n_i} f(y_{it} | \mathbf{d}_i, \beta, \phi) f(\mathbf{d}_i | \mathbf{G}) d(\mathbf{d}_i) \quad (30)$$

sendo que o número de integrais na expressão (30) será o número de termos em  $\mathbf{d}_i$  e  $f(\mathbf{d}_i | \mathbf{G})$  é a função densidade dos efeitos aleatórios.

No caso particular em que  $Y_{it} \sim N(\mathbf{X}_i \beta; \Sigma_i)$ , as integrais na equação (30) podem ser tratadas analiticamente. Em geral, isso não ocorre e, portanto, métodos numéricos devem ser utilizados para uma solução aproximada das integrais (VIEIRA, 2008).

O critério de verificação do ajuste recomendado por Littell et al. (2006) é a estatística Qui-Quadrado Generalizada, dada pela expressão  $Q'_g = \hat{\mathbf{r}}' [\mathbf{V}(\mathbf{Y}_i)]^{-1} \hat{\mathbf{r}}$ . Se a divisão dessa estatística pelo grau de liberdade do resíduo for próximo a 1, tem-se um bom ajuste do modelo.

## **5 MATERIAL E MÉTODOS**

### **5.1 Material**

O experimento foi realizado no período de novembro de 2005 a junho de 2006, no Laboratório de Sementes do Departamento de Produção Vegetal da Faculdade de Ciências Agrônômicas – UNESP, Fazenda Lageado, Botucatu-SP, pela engenheira agrônoma Líbia Belisário da Silva (SILVA, 2007).

O delineamento experimental foi fatorial inteiramente casualizado do tipo 9 x 4, sendo 9 épocas de colheita de sementes e 4 condições diferentes de armazenamento: sementes sem armazenamento (SA) (sementes extraídas no momento da colheita, sem repouso), condição nua (sementes extraídas e colocadas em repouso), condição fruto (sementes nos frutos soltos colocadas em repouso) e condição cacho (sementes nos frutos presos ao racemo colocadas em repouso).

As colheitas de sementes foram realizadas manualmente, com auxílio de tesoura de poda, em intervalos de 14 dias durante o período de 30 a 142 dias após a antese, e coletando-se sempre o primeiro cacho. Considerou-se como antese o dia em que as primeiras flores femininas das inflorescências estavam abertas. Assim, foram realizadas nove colheitas, iniciando-se a primeira no dia 20 de fevereiro de 2006 e a última no dia 5 de junho de 2006, conforme mostrado na tabela 3.

Tabela 3: Épocas de colheita dos racemos de mamoneira

Data das colheitas	Dias após a antese
20/02/2006	30
03/03/2006	44
09/03/2006	58
27/03/2006	72
06/04/2006	86
20/04/2006	100
08/05/2006	114
22/05/2006	128
05/06/2006	142

Em cada época foram coletados 30 racemos, os quais foram separados em três grupos, com 10 racemos cada. No primeiro grupo, as sementes foram extraídas dos frutos, no segundo grupo os frutos foram removidos dos racemos e no terceiro grupo, os racemos foram mantidos intactos.

O teste de germinação foi instalado com oito repetições de 25 sementes por tratamento, em rolos de papel toalha tipo Germitest dispostos na posição horizontal em germinador mantido em temperaturas alternadas de 20–30°C em papel umedecido com 2,5 vezes a massa (g) do papel em água. Durante os testes, os rolos de papel permaneceram acondicionados dentro de sacos plásticos de 0,033 mm de espessura fechados, para evitar a desidratação (SILVA, 2007). As avaliações foram realizadas aos 7, 14 e 21 dias após a semeadura, seguindo os critérios estabelecidos em Brasil (1992).

## 5.2 Métodos

Para a análise estatística desses dados, após uma análise descritiva e gráfica, considerou-se duas respostas: proporção de sementes germinadas no final do período e proporção de sementes germinadas em cada um dos três períodos. Essas proporções são condicionais e, portanto, não há a dependência ao longo do tempo.

Em cada tipo de resposta foram utilizados os três enfoques: o do modelo linear generalizado (MLG) onde tem-se somente efeitos fixos, o de modelos condicionais (MLGM), utilizando-se os modelos com efeitos fixos e com efeito aleatório de rolo de papel e o de modelos marginais (EEG), utilizando-se as equações de estimação generalizada. As análises foram feitas utilizando-se o sistema SAS – *Statistical Analysis System*, versão 9.2 (SAS INSTITUTE, 2008) e os gráficos foram feitos no sistema estatístico R, versão 2.10.1 (R DEVELOPMENT CORE TEAM, 2010).

A distribuição padrão a ser assumida para a variável aleatória  $Y$ , que é o número de sementes germinadas, é a binomial com parâmetro  $m$  e probabilidade de sucesso  $\pi$ , isto é,

$$Y \sim \text{Bin}(m; \pi).$$

Para o modelo que leva em consideração somente os efeitos fixos (MLG), utilizou-se as funções de ligação complemento log-log (cll) e logística (logit). Considerando a proporção de sementes germinadas no final do período e utilizando a função de ligação cll, tem-se o modelo complemento log-log (MLG1) dado na equação (31). Utilizando a função de ligação logit, tem-se o modelo logístico (MLG2) dado na equação (32).

$$\eta_{jk} = \text{cll}(\pi_{jk}) = \ln[-\ln(1 - \pi_{jk})] = \mu + \beta_j + \tau_k + (\beta\tau)_{jk} \quad (31)$$

$$\eta_{jk} = \text{logit}(\pi_{jk}) = \ln\left(\frac{\pi_{jk}}{1 - \pi_{jk}}\right) = \mu + \beta_j + \tau_k + (\beta\tau)_{jk} \quad (32)$$

onde:

$\mu$  é o efeito associado à média geral;

$\beta_j$  é o efeito associado à  $j$ -ésima condição de armazenamento,  $j = 1, 2, 3, 4$ ;

$\tau_k$  é o efeito associado à  $k$ -ésima época,  $k = 1, \dots, 9$ ;

$(\beta\tau)_{jk}$  é o efeito da interação condição de armazenamento e época.

Considerando-se a proporção de sementes germinadas em cada período (7, 14 e 21 dias) e utilizando a metodologia de modelo linear generalizado com função de ligação cll, tem-se o modelo complemento log-log (MLG3) dado na equação (33). Utilizando a função de ligação logit, tem-se o modelo logístico (MLG4), dado na equação (34).

$$\eta_{ijk} = cll(\pi_{ijk}) = \ln[-\ln(1 - \pi_{ijk})] = \mu + \alpha_i + \beta_j + \tau_k + (\alpha\beta)_{ij} + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} \quad (33)$$

$$\eta_{ijk} = \text{logit}(\pi_{ijk}) = \ln\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) = \mu + \alpha_i + \beta_j + \tau_k + (\alpha\beta)_{ij} + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} \quad (34)$$

onde:

$\mu$  é o efeito associado à média geral;

$\alpha_i$  é o efeito associado ao  $i$ -ésimo período,  $i = 1, 2, 3$ ;

$\beta_j$  é o efeito associado à  $j$ -ésima condição de armazenamento,  $j = 1, 2, 3, 4$ ;

$\tau_k$  é o efeito associado à  $k$ -ésima época,  $k = 1, \dots, 9$ ;

$(\alpha\beta)_{ij}$  é o efeito da interação período e condição de armazenamento;

$(\alpha\tau)_{ik}$  é o efeito da interação período e época;

$(\beta\tau)_{jk}$  é o efeito da interação condição de armazenamento e época;

$(\alpha\beta\tau)_{ijk}$  é o efeito da interação período, condição de armazenamento e época.

A verificação do ajuste dos modelos lineares generalizados foi efetuada pelos critérios de *Deviance*, Qui-Quadrado de Pearson e AIC e também pelos gráficos dos resíduos de Pearson e probabilidade normal.

Para o modelo linear generalizado misto (MLGM), que é o modelo que leva em consideração os efeitos fixos e o efeito aleatório de rolo de papel, utilizou-se as funções de ligação complemento log-log e logística.

Considerando-se a proporção de sementes germinadas no final do período, o preditor linear para o modelo linear generalizado misto com função de ligação complemento log-log (MLGM1) é dado na equação (35), e com função de ligação logística (MLGM2) é dado na equação (36).

$$\eta_{jkl} = cll(\pi_{jkl}) = \ln[-\ln(1 - \pi_{jkl})] = \mu + \beta_j + \tau_k + (\beta\tau)_{jk} + d_{jkl} \quad (35)$$

$$\eta_{jkl} = \text{logit}(\pi_{jkl}) = \ln\left(\frac{\pi_{jkl}}{1 - \pi_{jkl}}\right) = \mu + \beta_j + \tau_k + (\beta\tau)_{jk} + d_{jkl} \quad (36)$$

onde:

$\mu$  é o efeito associado à média geral;

$\beta_j$  é o efeito associado à  $j$ -ésima condição de armazenamento,  $j = 1, 2, 3, 4$ ;

$\tau_k$  é o efeito associado à  $k$ -ésima época,  $k = 1, \dots, 9$ ;

$(\beta\tau)_{jk}$  é o efeito da interação condição de armazenamento e época;

$d_{jkl} \sim$  é o efeito aleatório de rolo de papel com  $N(0; \sigma^2)$ ,  $l = 1, \dots, 8$ .

Considerando a proporção de sementes germinadas em cada período (7, 14 e 21 dias) e utilizando a metodologia de modelo linear generalizado misto com função de ligação cll, tem-se o modelo linear generalizado misto complemento log-log (MLGM3) dado na equação (37). Utilizando a função de ligação logit, tem-se o modelo linear generalizado misto logístico (MLGM4) dado na equação (38).

$$\eta_{ijkl} = \ln[-\ln(1 - \pi_{ijkl})] = \mu + \alpha_i + \beta_j + \tau_k + (\alpha\beta)_{ij} + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} + d_{ijkl} \quad (37)$$

$$\eta_{ijkl} = \ln\left(\frac{\pi_{ijkl}}{1 - \pi_{ijkl}}\right) = \mu + \alpha_i + \beta_j + \tau_k + (\alpha\beta)_{ij} + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} + d_{ijkl} \quad (38)$$

onde:

$\mu$  é o efeito associado à média geral;

$\alpha_i$  é o efeito associado ao  $i$ -ésimo período,  $i = 1, 2, 3$ ;

$\beta_j$  é o efeito associado à  $j$ -ésima condição de armazenamento,  $j = 1, 2, 3, 4$ ;

$\tau_k$  é o efeito associado à  $k$ -ésima época,  $k = 1, \dots, 9$ ;

$(\alpha\beta)_{ij}$  é o efeito da interação período e condição de armazenamento;

$(\alpha\tau)_{ik}$  é o efeito da interação período e época;

$(\beta\tau)_{jk}$  é o efeito da interação condição de armazenamento e época;

$(\alpha\beta\tau)_{ijk}$  é o efeito da interação período, condição de armazenamento e época;

$d_{ijkl} \sim$  é o efeito aleatório de rolo de papel com  $N(0; \sigma^2)$ ,  $l = 1, \dots, 8$ .

A verificação do ajuste dos modelos lineares generalizados mistos foi efetuada pelo critério Qui-Quadrado Generalizado e também pelos gráficos dos resíduos de Pearson condicional e probabilidade normal.

Para o método da equação de estimação generalizada (EEG) que leva em consideração uma matriz de correlação de trabalho, utilizou-se as funções de ligação complemento log-log e logística.

Considerando a proporção de sementes germinadas no final do período, o preditor linear para a equação de estimação generalizada com função de ligação

complemento log-log (EEG1) é dado na equação (39) e com função de ligação logística (EEG2) é dado na equação (40).

$$\eta_{jk} = cll(\pi_{jk}) = \ln[-\ln(1 - \pi_{jk})] = \mu + \beta_j + \tau_k + (\beta\tau)_{jk} \quad (39)$$

$$\eta_{jk} = \text{logit}(\pi_{jk}) = \ln\left(\frac{\pi_{jk}}{1 - \pi_{jk}}\right) = \mu + \beta_j + \tau_k + (\beta\tau)_{jk} \quad (40)$$

onde:

$\mu$  é o efeito associado à média geral;

$\beta_j$  é o efeito associado à  $j$ -ésima condição de armazenamento,  $j = 1, 2, 3, 4$ ;

$\tau_k$  é o efeito associado à  $k$ -ésima época,  $k = 1, \dots, 9$ ;

$(\beta\tau)_{jk}$  é o efeito da interação condição de armazenamento e época.

No ajuste do modelo marginal através do método da equação de estimação generalizada (EEG), considerando a proporção de sementes germinadas em cada período (7, 14 e 21 dias) e utilizando a função de ligação complemento log-log, tem-se o modelo da equação de estimação generalizada complemento log-log (EEG3) dado na equação (41). Utilizando a função de ligação logística, tem-se o modelo da equação de estimação generalizada logístico (EEG4) dado na equação (42).

$$\eta_{ijk} = cll(\pi_{ijk}) = \ln[-\ln(1 - \pi_{ijk})] = \mu + \alpha_i + \beta_j + \tau_k + (\alpha\beta)_{ij} + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} \quad (41)$$

$$\eta_{ijk} = \text{logit}(\pi_{ijk}) = \ln\left(\frac{\pi_{ijk}}{1 - \pi_{ijk}}\right) = \mu + \alpha_i + \beta_j + \tau_k + (\alpha\beta)_{ij} + (\alpha\tau)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk} \quad (42)$$



onde:

$\mu$  é o efeito associado à média geral;

$\alpha_i$  é o efeito associado ao  $i$ -ésimo período,  $i = 1, 2, 3$ ;

$\beta_j$  é o efeito associado à  $j$ -ésima condição de armazenamento,  $j = 1, 2, 3, 4$ ;

$\tau_k$  é o efeito associado à  $k$ -ésima época,  $k = 1, \dots, 9$ ;

$(\alpha\beta)_{ij}$  é o efeito da interação período e condição de armazenamento;

$(\alpha\tau)_{ik}$  é o efeito da interação período e época;

$(\beta\tau)_{jk}$  é o efeito da interação condição de armazenamento e época;

$(\alpha\beta\tau)_{ijk}$  é o efeito da interação período, condição de armazenamento e época.

O processo de estimação dos parâmetros é semelhante ao caso em que se supõem observações independentes, sendo que uma matriz de correlação de trabalho é introduzida para se levar em consideração a correlação entre as 25 sementes do mesmo rolo de papel.

A verificação do ajuste das equações de estimação generalizadas foi efetuada pelo critério Qui-Quadrado Generalizado e também pelos gráficos dos resíduos de Pearson marginal e probabilidade normal.

## **6 RESULTADOS E DISCUSSÃO**

### **6.1 Análise descritiva dos dados pelo método de Kaplan-Meier**

O método Kaplan-Meier gera uma função escada que salta em cada tempo onde ocorreu um evento. O tamanho desse salto depende do número de eventos observados nesse tempo e também do número de observações censuradas antes dele. Essa função é em escada porque se assume que o risco mantém-se constante até a ocorrência do próximo evento.

Na tabela 4, tem-se que o maior número de sementes germinadas ocorreu na época 86, com 686 sementes; e somente 6 sementes germinaram na época 30. Nas épocas de colheita 44, 58, 72, 100, 114, 128 e 142, os números de sementes germinadas foram, respectivamente, 130, 388, 609, 665, 636, 628 e 494.

Como somente poucas sementes germinaram no início do experimento, retirou-se a época 30 das análises, pois sua inclusão em algumas análises faz com que o algoritmo de estimação não tenha convergência.

Na figura 6, observa-se o comportamento de cada época em cada tratamento. O tratamento sem armazenamento (SA) apresentou 1087 sementes germinadas, o tratamento cacho apresentou 1100 sementes germinadas, o tratamento nua apresentou 1048 sementes germinadas e o tratamento fruto apresentou 1007 sementes germinadas.

Tabela 4: Estimativas da função de sobrevivência pelo método de Kaplan-Meier

Época	$t_i$	$R(t_i)$	$N(t_i)$	$\hat{S}(t)$	Erro padrão	I.C. 95%
30	7	2399	0	1,000	---	---
	14	1599	1	0,999	0,0006	[0,998; 1,000]
	21	799	5	0,993	0,0029	[0,988; 0,999]
44	7	2248	30	0,987	0,0024	[0,982; 0,991]
	14	1448	92	0,924	0,0067	[0,911; 0,937]
	21	678	8	0,913	0,0077	[0,898; 0,928]
58	7	2077	32	0,985	0,0027	[0,979; 0,990]
	14	1277	259	0,785	0,0113	[0,763; 0,807]
	21	509	97	0,635	0,0164	[0,604; 0,668]
72	7	1855	47	0,975	0,0037	[0,968; 0,982]
	14	1055	451	0,558	0,0150	[0,529; 0,588]
	21	302	111	0,353	0,0182	[0,319; 0,390]
86	7	1626	117	0,928	0,0064	[0,916; 0,941]
	14	826	540	0,321	0,0155	[0,292; 0,353]
	21	143	29	0,256	0,0164	[0,226; 0,290]
100	7	1625	149	0,908	0,0072	[0,894; 0,922]
	14	825	477	0,383	0,0159	[0,353; 0,416]
	21	174	39	0,297	0,0173	[0,265; 0,333]
114	7	1474	333	0,774	0,0109	[0,753; 0,796]
	14	674	260	0,475	0,0160	[0,445; 0,508]
	21	207	43	0,377	0,0184	[0,342; 0,415]
128	7	1599	207	0,871	0,0084	[0,854; 0,887]
	14	801	385	0,452	0,0160	[0,422; 0,485]
	21	208	36	0,374	0,0178	[0,341; 0,410]
142	7	1848	79	0,957	0,0047	[0,948; 0,967]
	14	1048	394	0,597	0,0146	[0,569; 0,627]
	21	327	21	0,559	0,0159	[0,529; 0,591]

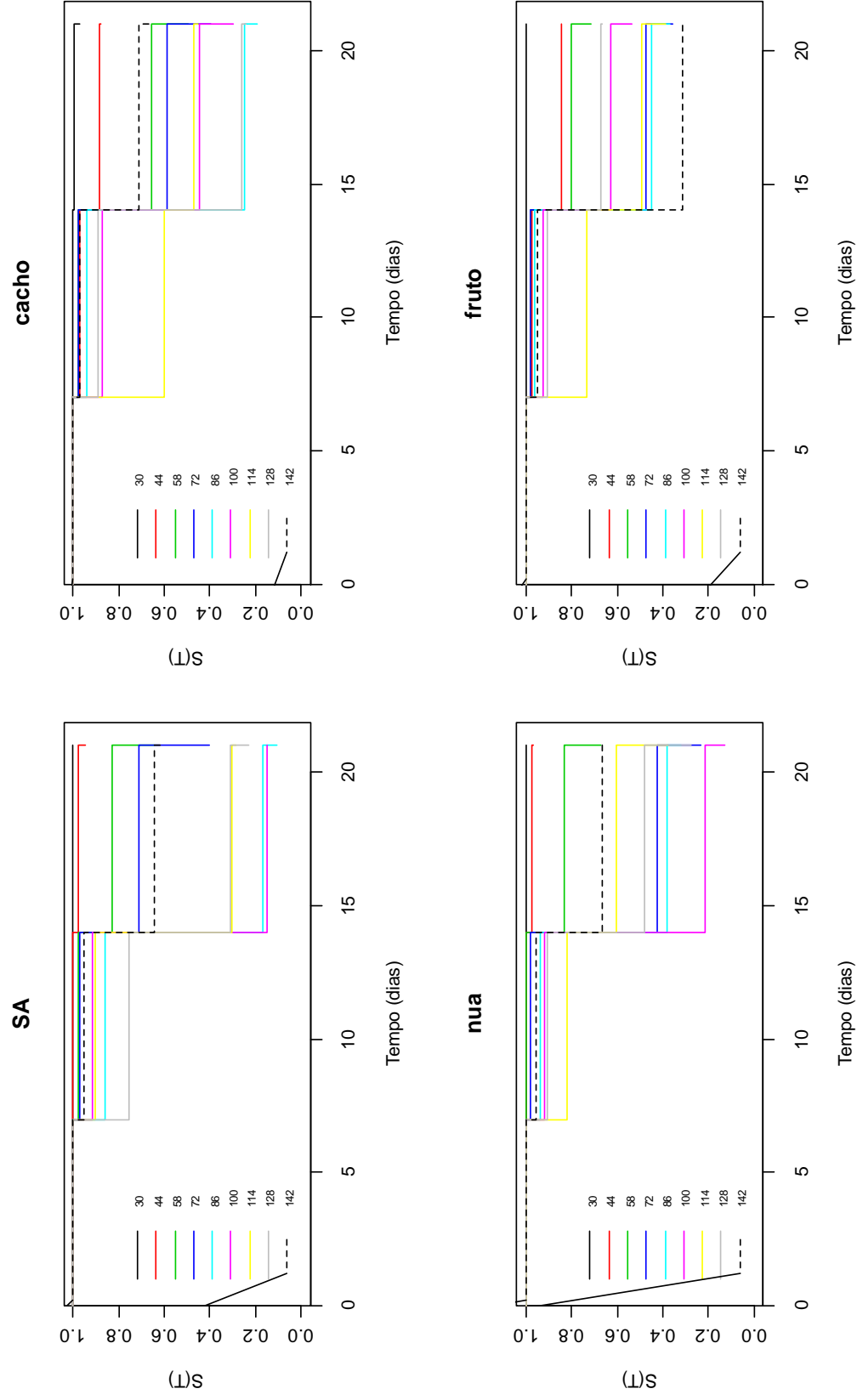


Figura 6: Gráficos Kaplan-Meier para cada tratamento, de acordo com as épocas de colheita.

Na tabela 5, observa-se que no período 7 tem-se 994 sementes germinadas, sendo 294 no tratamento cacho, 241 no tratamento fruto, 203 no tratamento nua e 256 no tratamento SA. No período 14 tem-se 2859 sementes germinadas, sendo 716 no tratamento cacho, 702 no tratamento fruto, 717 no tratamento nua e 724 no tratamento SA. No período 21 tem-se 389 sementes germinadas, sendo 90 no tratamento cacho, 64 no tratamento fruto, 128 no tratamento nua e 107 no tratamento SA.

Tabela 5: Estimativas da função de sobrevivência por Kaplan-Meier

Tratamento	$t_i$	$R(t_i)$	$N(t_i)$	$\hat{S}(t)$	Erro padrão	I.C. 95%
cacho	7	4095	294	0,928	0,0040	[0,920; 0,936]
	14	2296	716	0,639	0,0094	[0,621; 0,657]
	21	790	90	0,566	0,0110	[0,545; 0,588]
fruto	7	4216	241	0,943	0,0036	[0,936; 0,950]
	14	2416	702	0,669	0,0091	[0,651; 0,687]
	21	857	64	0,619	0,0103	[0,599; 0,639]
nua	7	4277	203	0,953	0,0033	[0,946; 0,959]
	14	2477	717	0,677	0,0090	[0,659; 0,695]
	21	880	128	0,578	0,0111	[0,557; 0,601]
SA	7	4163	256	0,939	0,0037	[0,931; 0,946]
	14	2364	724	0,651	0,0093	[0,633; 0,669]
	21	820	107	0,566	0,0111	[0,545; 0,588]

O Kaplan-Meier é visto como uma análise exploratória dos dados. Para uma análise mais formal, consideram-se duas respostas: proporção de sementes germinadas no final do experimento e proporção de sementes germinadas em cada momento.

## 6.2 Análise considerando a proporção de sementes germinadas no final do período

Utilizando o modelo linear generalizado (MLG), visto na seção 4.4, com função de ligação complemento log-log (MLG1) e logística (MLG2), obtiveram-se as estimativas dos parâmetros e seus erros-padrão, apresentados na tabela 6.

Tabela 6: Resultados do ajuste do MLG1 e MLG2

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLG1	MLG2	MLG1	MLG2	MLG1	MLG2
$\mu$	-0,3665	0,0000	0,1020	0,1414	0,0003	1,0000
$\beta_1$	0,2654	0,3847	0,1393	0,2019	0,0567	0,0567
$\beta_2$	0,1415	0,2007	0,1414	0,2005	0,3169	0,3169
$\beta_3$	0,9221	1,5506	0,1348	0,2337	< 0,0001	< 0,0001
$\tau_1$	-0,9753	-1,2083	0,1796	0,2196	< 0,0001	< 0,0001
$\tau_2$	0,3879	0,5754	0,1376	0,2042	0,0048	0,0048
$\tau_3$	0,6219	0,9694	0,1353	0,2123	< 0,0001	< 0,0001
$\tau_4$	1,1791	2,1429	0,1371	0,2706	< 0,0001	< 0,0001
$\tau_5$	0,9386	1,5856	0,1348	0,2354	< 0,0001	< 0,0001
$\tau_6$	0,7664	1,2368	0,1347	0,2206	< 0,0001	< 0,0001
$\tau_7$	1,0986	1,9459	0,1360	0,2563	< 0,0001	< 0,0001
$(\beta\tau)_{11}$	-1,4751	-1,6887	0,3286	0,3756	< 0,0001	< 0,0001
$(\beta\tau)_{21}$	-1,6719	-1,8362	0,3644	0,4058	< 0,0001	< 0,0001
$(\beta\tau)_{31}$	-0,6518	-1,2377	0,2397	0,3273	0,0065	0,0002
$(\beta\tau)_{12}$	-0,7711	-1,1204	0,1978	0,2874	< 0,0001	< 0,0001
$(\beta\tau)_{22}$	-0,7546	-1,0783	0,2014	0,2870	0,0002	0,0002
$(\beta\tau)_{32}$	-1,5669	-2,4693	0,1975	0,3113	< 0,0001	< 0,0001
$(\beta\tau)_{13}$	-0,3490	-0,5305	0,1881	0,2990	0,0635	0,0761
$(\beta\tau)_{23}$	0,2611	0,6043	0,1892	0,3250	0,1676	0,0630
$(\beta\tau)_{33}$	-0,7925	-1,3117	0,1839	0,3285	< 0,0001	< 0,0001
$(\beta\tau)_{14}$	0,0191	0,4169	0,1959	0,4463	0,9222	0,3503
$(\beta\tau)_{24}$	-0,4309	-0,8607	0,1900	0,3558	0,0234	0,0156

Tabela 6: Resultados do ajuste do MLG1 e MLG2

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLG1	MLG2	MLG1	MLG2	MLG1	MLG2
$(\beta\tau)_{34}$	-1,3497	-2,4851	0,1852	0,3689	< 0,0001	< 0,0001
$(\beta\tau)_{15}$	0,1142	0,5420	0,1907	0,3850	0,5491	0,1592
$(\beta\tau)_{25}$	0,2919	0,8799	0,1934	0,3974	0,1311	0,0268
$(\beta\tau)_{35}$	-1,4864	-2,5825	0,1858	0,3341	< 0,0001	< 0,0001
$(\beta\tau)_{16}$	-0,0424	0,0745	0,1870	0,3280	0,8206	0,8204
$(\beta\tau)_{26}$	-0,1563	-0,2291	0,1884	0,3116	0,4065	0,4622
$(\beta\tau)_{36}$	-0,8769	-1,4624	0,1834	0,3368	< 0,0001	< 0,0001
$(\beta\tau)_{17}$	-0,2058	-0,2398	0,1892	0,3709	0,2767	0,5178
$(\beta\tau)_{27}$	-0,2848	-0,5251	0,1894	0,3496	0,1327	0,1331
$(\beta\tau)_{37}$	-1,9071	-3,3362	0,1897	0,3471	< 0,0001	< 0,0001

Nas figuras 7 e 8 têm-se os gráficos dos resíduos de Pearson versus a ordem das observações para o MLG1 e MLG2, respectivamente.

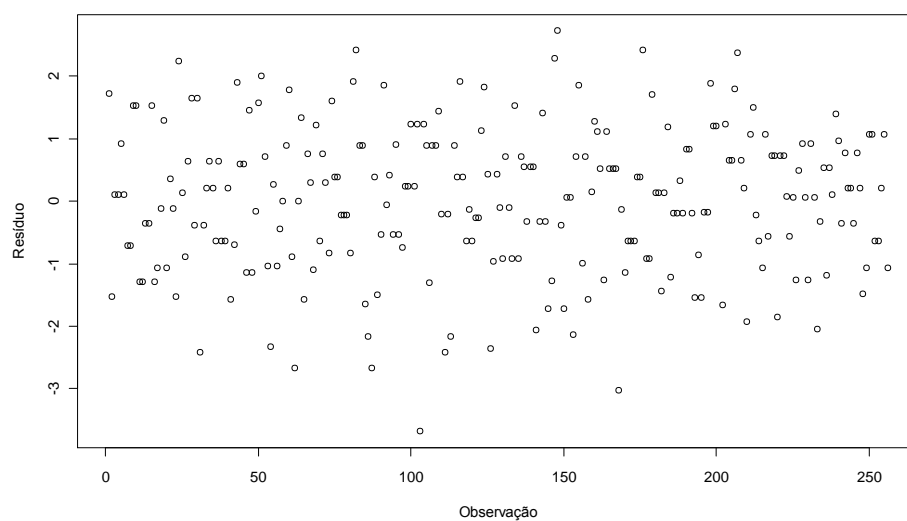


Figura 7: Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG1.

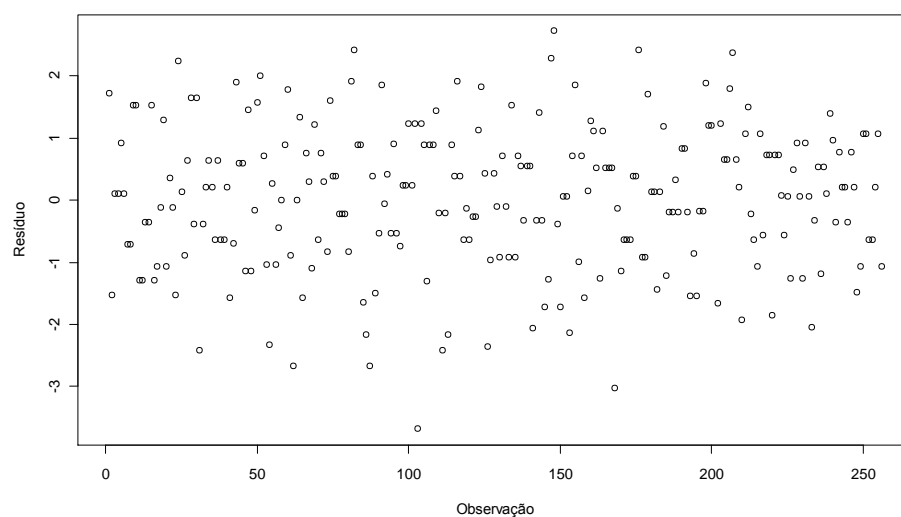


Figura 8: Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG2.

Nas figuras 9 e 10 têm-se os gráficos de probabilidade normal para os resíduos de Pearson do MLG1 e MLG2, respectivamente.

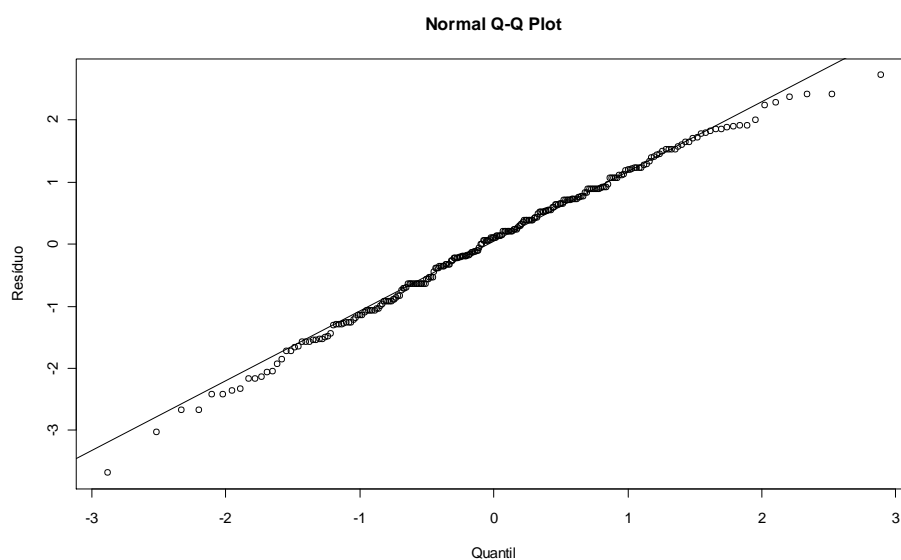


Figura 9: Gráfico de probabilidade normal para os resíduos de Pearson do MLG1.



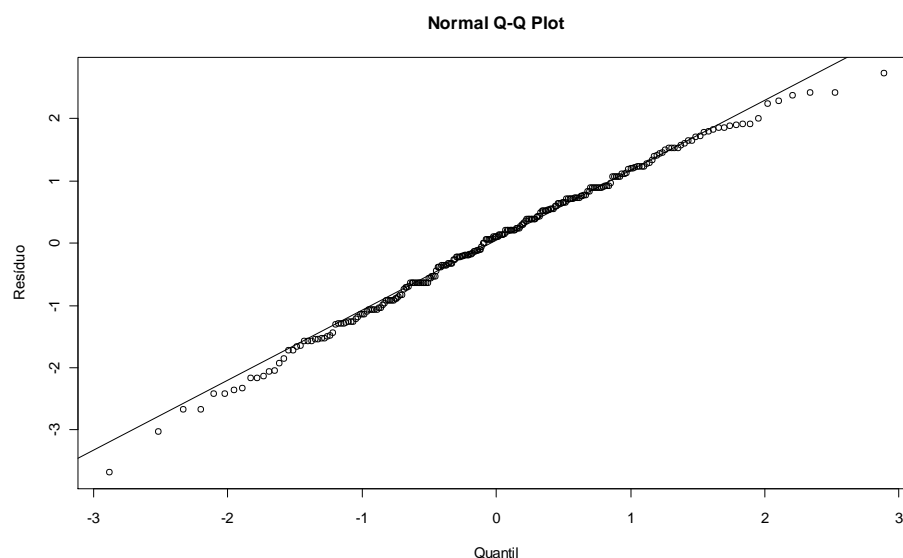


Figura 10: Gráfico de probabilidade normal para os resíduos de Pearson do MLG2.

Na tabela 7, têm-se os critérios de verificação do ajuste para o MLG1 com função de ligação complemento log-log e para o MLG2 com função de ligação logística.

Tabela 7: Critérios para verificação do ajuste do MLG1 e MLG2

Modelo	Critério	G.L.	Valor	Valor / G.L.
MLG1	<i>Deviance</i>	224	332,6570	1,4851
	Qui-Quadrado de Pearson	224	319,0575	1,4244
	AIC	—	1171,5281	—
MLG2	<i>Deviance</i>	224	332,6570	1,4851
	Qui-Quadrado de Pearson	224	319,0576	1,4244
	AIC	—	1171,5281	—

Os gráficos dos resíduos de Pearson (Figuras 7 e 8) e os gráficos de probabilidade normal (Figuras 9 e 10) apresentaram valores muito próximos. Os critérios de verificação do ajuste (Tabela 7) para as funções de ligação complemento log-log (MLG1) e logística (MLG2) apresentaram valores altos e próximos. Os valores dos resíduos padronizados estão na faixa de -4 a 3, indicando que os valores estão altos.

Utilizando o modelo linear generalizado misto (MLGM) com efeito aleatório de rolo de papel, visto na seção 4.6, com função de ligação complemento log-log (MLGM1) e logística (MLGM2), obtiveram-se os resultados apresentados na tabela 8.

Tabela 8: Resultados do ajuste do MLGM1 e MLGM2

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLGM1	MLGM2	MLGM1	MLGM2	MLGM1	MLGM2
$\mu$	-0,3665	0,0000	0,1128	0,1776	0,0013	0,9999
$\beta_1$	0,2654	0,3857	0,1550	0,2526	0,0882	0,1282
$\beta_2$	0,1415	0,2015	0,1569	0,2516	0,3681	0,4239
$\beta_3$	0,9221	1,5528	0,1510	0,2787	< 0,0001	< 0,0001
$\tau_1$	-0,9753	-1,2155	0,1921	0,2672	< 0,0001	< 0,0001
$\tau_2$	0,3879	0,5795	0,1535	0,2547	0,0122	0,0238
$\tau_3$	0,6219	0,9731	0,1515	0,2611	< 0,0001	0,0002
$\tau_4$	1,1791	2,1496	0,1531	0,3105	< 0,0001	< 0,0001
$\tau_5$	0,9386	1,5942	0,1510	0,2805	< 0,0001	< 0,0001
$\tau_6$	0,7664	1,2385	0,1509	0,2678	< 0,0001	< 0,0001
$\tau_7$	1,0986	1,9496	0,1521	0,2980	< 0,0001	< 0,0001
$(\beta\tau)_{11}$	-1,4751	-1,6861	0,3424	0,4328	< 0,0001	0,0001
$(\beta\tau)_{21}$	-1,6719	-1,8351	0,3769	0,4593	< 0,0001	< 0,0001
$(\beta\tau)_{31}$	-0,6518	-1,2378	0,2583	0,3917	0,0123	0,0018
$(\beta\tau)_{12}$	-0,7711	-1,1258	0,2200	0,3588	0,0006	0,0019
$(\beta\tau)_{22}$	-0,7546	-1,0853	0,2232	0,3587	0,0009	0,0028
$(\beta\tau)_{32}$	-1,5669	-2,4787	0,2197	0,3785	< 0,0001	< 0,0001
$(\beta\tau)_{13}$	-0,3490	-0,5323	0,2113	0,3682	0,1000	0,1496
$(\beta\tau)_{23}$	0,2611	0,6024	0,2123	0,3896	0,2200	0,1235
$(\beta\tau)_{33}$	-0,7925	-1,3035	0,2076	0,3929	0,0002	0,0011

Tabela 8: Resultados do ajuste do MLGM1 e MLGM2

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLGM1	MLGM2	MLGM1	MLGM2	MLGM1	MLGM2
$(\beta\tau)_{14}$	0,0191	0,4169	0,2183	0,4956	0,9302	0,4011
$(\beta\tau)_{24}$	-0,4309	-0,8611	0,2130	0,4159	0,0443	0,0396
$(\beta\tau)_{34}$	-1,3497	-2,4885	0,2087	0,4271	< 0,0001	< 0,0001
$(\beta\tau)_{15}$	0,1142	0,5353	0,2136	0,4411	0,5933	0,2261
$(\beta\tau)_{25}$	0,2919	0,8739	0,2160	0,4519	0,1779	0,0544
$(\beta\tau)_{35}$	-1,4864	-2,5866	0,2092	0,3976	< 0,0001	< 0,0001
$(\beta\tau)_{16}$	-0,0424	0,0803	0,2103	0,3922	0,8403	0,8379
$(\beta\tau)_{26}$	-0,1563	-0,2268	0,2115	0,3785	0,4606	0,5497
$(\beta\tau)_{36}$	-0,8769	-1,4617	0,2071	0,3995	< 0,0001	0,0003
$(\beta\tau)_{17}$	-0,2058	-0,2377	0,2123	0,4287	0,3334	0,5799
$(\beta\tau)_{27}$	-0,2848	-0,5022	0,2125	0,4113	0,1814	0,2234
$(\beta\tau)_{37}$	-1,9071	-3,3413	0,2127	0,4082	< 0,0001	< 0,0001

Nas figuras 11 e 12 têm-se os gráficos dos resíduos de Pearson condicional versus a ordem das observações para o MLGM1 e MLGM2, respectivamente.

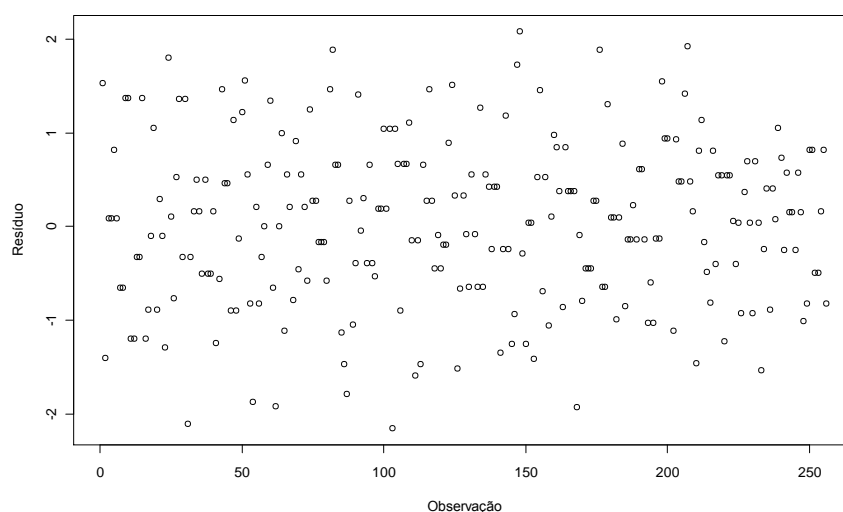


Figura 11: Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM1

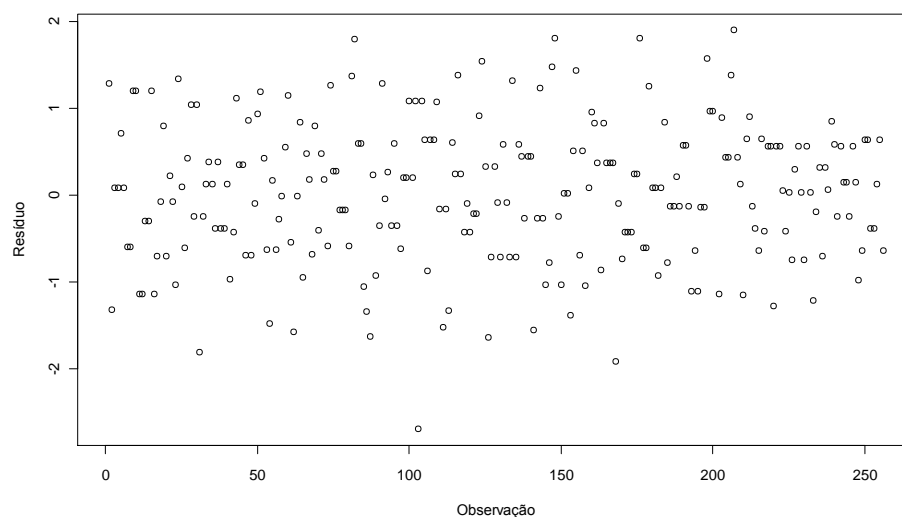


Figura 12: Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM2.

Nas figuras 13 e 14 têm-se os gráficos de probabilidade normal para os resíduos de Pearson condicional do MLGM1 e MLGM2, respectivamente.

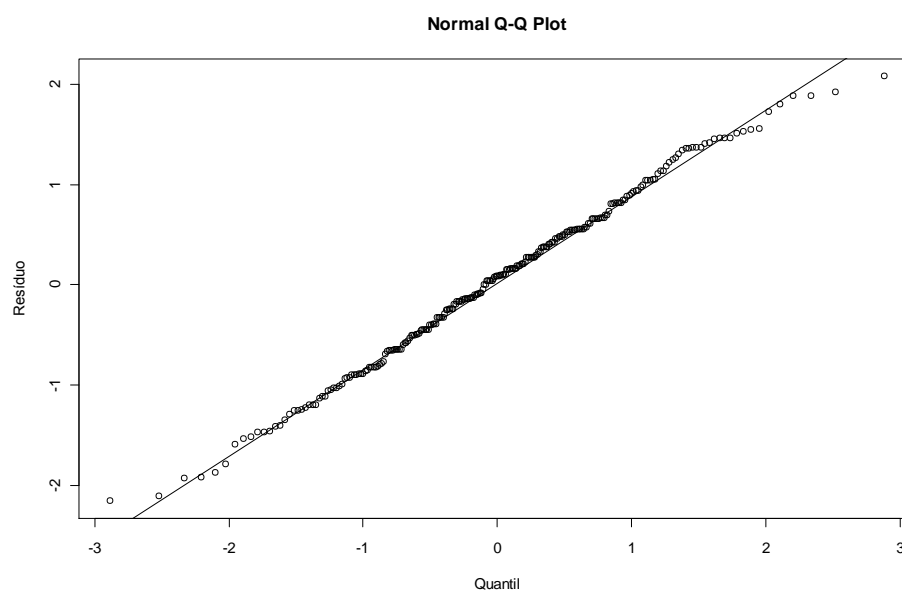


Figura 13: Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM1.

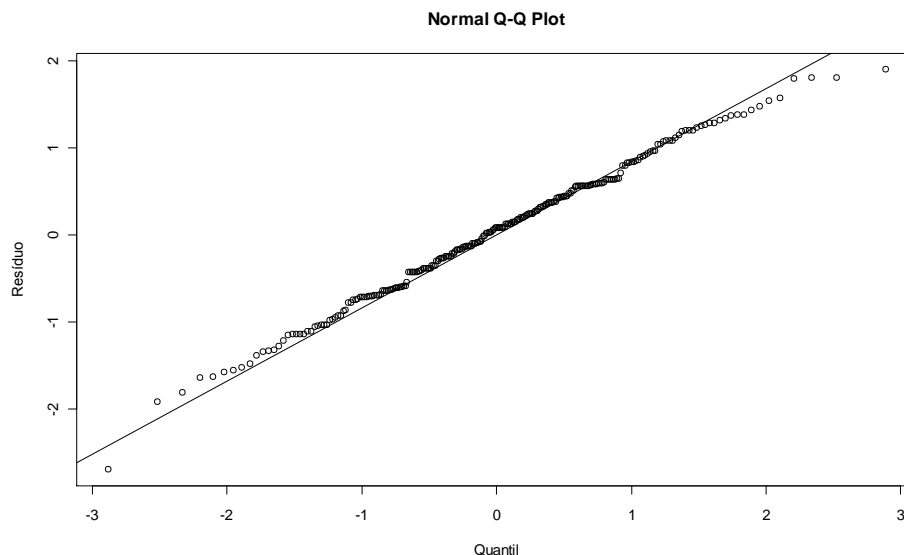


Figura 14: Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM2.

Comparando-se os gráficos dos resíduos de Pearson e de probabilidade normal do MLG1 e MLG2 com os gráficos do MLGM1 e MLGM2 verifica-se uma melhora no ajuste dos modelos com efeito aleatório, com variação menor nos valores dos resíduos.

No modelo linear generalizado misto, a qualidade do ajuste é verificada pelo cálculo do Qui-Quadrado Generalizado. Quanto mais próximo de 1, melhor o ajuste do modelo aos dados. Na tabela 9 observa-se que o modelo linear generalizado misto, que considera a função de ligação logística (MLGM2), ajustou-se melhor aos dados ( $Q'_g = 1,0192$ ).

Tabela 9: Critério para verificação do ajuste do MLGM1 e MLGM2

Modelo	Critério	G.L.	Valor	Valor / G.L.
MLGM1	Qui-Quadrado Generalizado	224	256,64	1,1457
MLGM2	Qui-Quadrado Generalizado	224	228,31	1,0192

Utilizando equação de estimação generalizada (EEG), visto na seção 4.5, com matriz de correlação não-estruturada e função de ligação complemento log-log (EEG1) e logística (EEG2), obtiveram-se os resultados apresentados na tabela 10. As estimativas do EEG1 e EEG2 foram iguais às estimativas do MLG1 e MLG2, respectivamente, mas os erros-padrão do EEG foram menores.

Tabela 10: Resultados do ajuste do EEG1 e EEG2

Parâmetro	Estimativa		Erro-padrão		Valor p	
	EEG1	EEG2	EEG1	EEG2	EEG1	EEG2
$\mu$	-0,3665	0,0000	0,1218	0,1688	0,0029	1,0000
$\beta_1$	0,2654	0,3847	0,1662	0,2409	0,1117	0,1117
$\beta_2$	0,1415	0,2007	0,1687	0,2393	0,4026	0,4026
$\beta_3$	0,9221	1,5506	0,1609	0,2790	< 0,0001	< 0,0001
$\tau_1$	-0,9753	-1,2083	0,2144	0,2621	< 0,0001	< 0,0001
$\tau_2$	0,3879	0,5754	0,1642	0,2437	0,0190	0,0191
$\tau_3$	0,6219	0,9694	0,1615	0,2534	0,0002	0,0002
$\tau_4$	1,1791	2,1429	0,1636	0,3229	< 0,0001	< 0,0001
$\tau_5$	0,9386	1,5856	0,1609	0,2810	< 0,0001	< 0,0001
$\tau_6$	0,7664	1,2368	0,1608	0,2633	< 0,0001	< 0,0001
$\tau_7$	1,0986	1,9459	0,1623	0,3059	< 0,0001	< 0,0001
$(\beta\tau)_{11}$	-1,4751	-1,6887	0,3921	0,4482	0,0002	0,0002
$(\beta\tau)_{21}$	-1,6719	-1,8362	0,4349	0,4843	0,0002	0,0002
$(\beta\tau)_{31}$	-0,6518	-1,2377	0,2860	0,3907	0,0236	0,0017
$(\beta\tau)_{12}$	-0,7711	-1,1204	0,2361	0,3430	0,0013	0,0013
$(\beta\tau)_{22}$	-0,7546	-1,0783	0,2404	0,3425	0,0019	0,0019
$(\beta\tau)_{32}$	-1,5669	-2,4693	0,2358	0,3716	< 0,0001	< 0,0001
$(\beta\tau)_{13}$	-0,3490	-0,5305	0,2245	0,3569	0,1214	0,1386
$(\beta\tau)_{23}$	0,2611	0,6043	0,2258	0,3879	0,2488	0,1206
$(\beta\tau)_{33}$	-0,7925	-1,3117	0,2195	0,3921	0,0004	0,0010

Tabela 10: Resultados do ajuste do EEG1 e EEG2

Parâmetro	Estimativa		Erro-padrão		Valor p	
	EEG1	EEG2	EEG1	EEG2	EEG1	EEG2
$(\beta\tau)_{14}$	0,0191	0,4169	0,2338	0,5327	0,9349	0,4347
$(\beta\tau)_{24}$	-0,4309	-0,8607	0,2268	0,4246	0,0587	0,0438
$(\beta\tau)_{34}$	-1,3497	-2,4851	0,2211	0,4402	< 0,0001	0,4402
$(\beta\tau)_{15}$	0,1142	0,5420	0,2276	0,4595	0,6162	0,2395
$(\beta\tau)_{25}$	0,2919	0,8799	0,2308	0,4743	0,2072	0,0649
$(\beta\tau)_{35}$	-1,4864	-2,5825	0,2217	0,3988	< 0,0001	< 0,0001
$(\beta\tau)_{16}$	-0,0424	0,0745	0,2232	0,3915	0,8494	0,8493
$(\beta\tau)_{26}$	-0,1563	-0,2291	0,2248	0,3719	0,4875	0,5385
$(\beta\tau)_{36}$	-0,8769	-1,4624	0,2189	0,4020	< 0,0001	0,0003
$(\beta\tau)_{17}$	-0,2058	-0,2398	0,2258	0,4426	0,3631	0,5884
$(\beta\tau)_{27}$	-0,2848	-0,5251	0,2261	0,4172	0,2090	0,2095
$(\beta\tau)_{37}$	-1,9071	-3,3362	0,2264	0,4142	< 0,0001	< 0,0001

Nas figuras 15 e 16 têm-se os gráficos dos resíduos de Pearson marginal versus a ordem das observações para o EEG1 e EEG2, respectivamente.

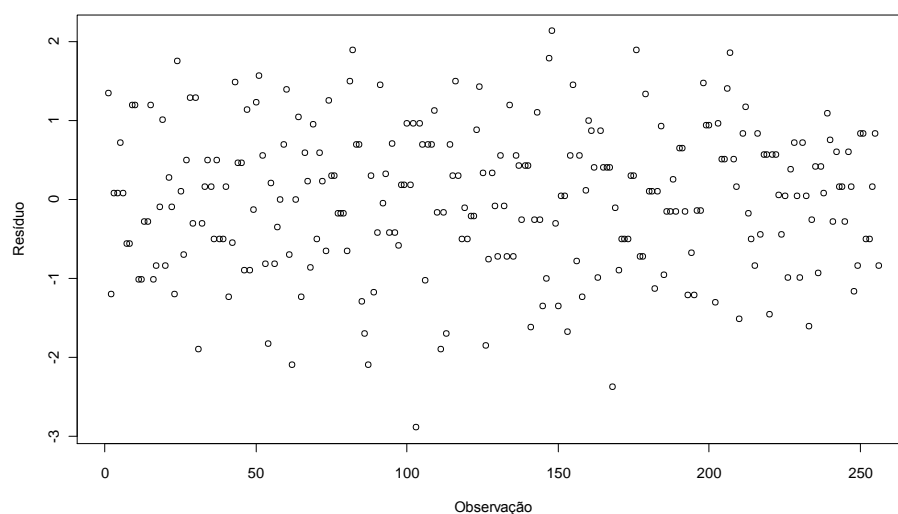


Figura 15: Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG1.

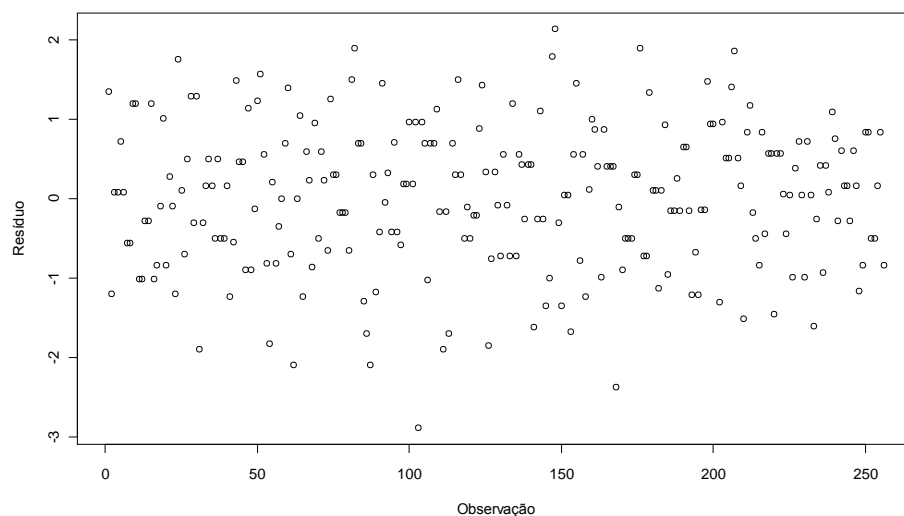


Figura 16: Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG2.

Nas figuras 17 e 18 têm-se os gráficos de probabilidade normal para os resíduos de Pearson marginal do EEG1 e EEG2, respectivamente.

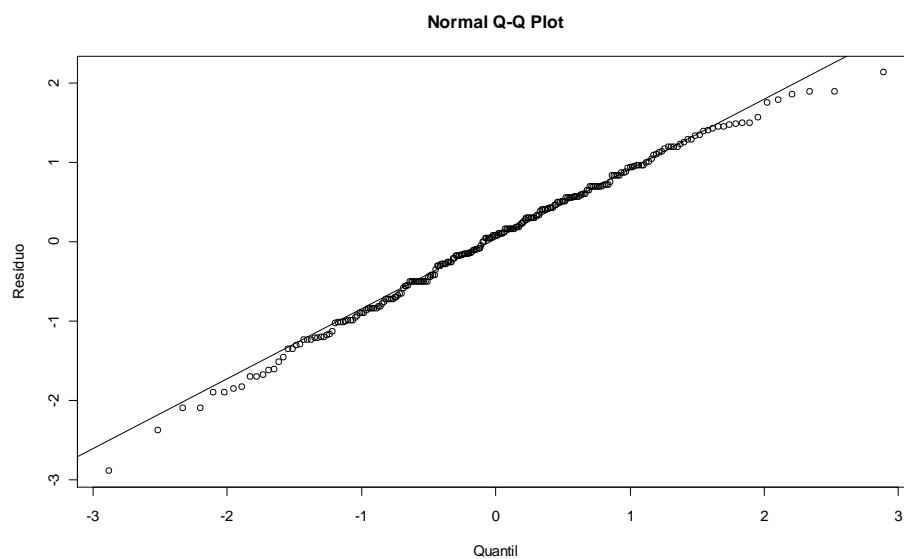


Figura 17: Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG1.



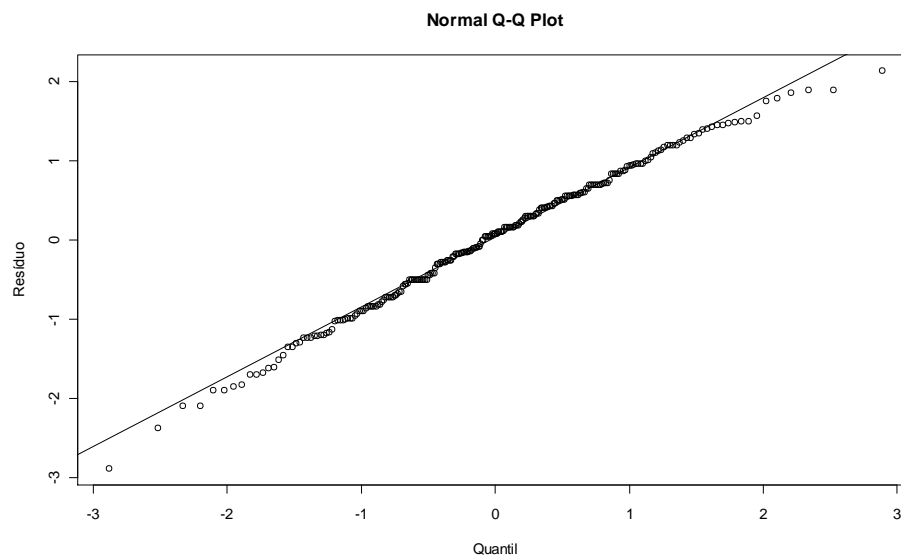


Figura 18: Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG2.

Os gráficos dos resíduos de Pearson marginal (Figuras 15 e 16) e os gráficos de probabilidade normal (Figuras 17 e 18) deram valores muito próximos.

Na equação de estimação generalizada, a qualidade do ajuste é verificada pelo cálculo do Qui-Quadrado Generalizado. Quanto mais próximo de 1, melhor o ajuste do modelo aos dados. Na tabela 11, observa-se que os dois modelos ajustaram-se satisfatoriamente ao conjunto de dados ( $Q_g = 1,0000$ ).

Tabela 11: Critério para verificação do ajuste do EEG1 e EEG2

Modelo	Critério	G.L.	Valor	Valor / G.L.
EEG1	Qui-Quadrado Generalizado	224	224,00	1,0000
EEG2	Qui-Quadrado Generalizado	224	224,00	1,0000

Comparando a faixa de variação dos valores dos resíduos das três metodologias, observa-se uma variação maior para o modelo linear generalizado e variações menores para o modelo linear generalizado misto e para a equação de estimação generalizada.

O critério de verificação do ajuste do MLGM foi próximo de 1, enquanto que no EEG foi igual a 1.

Portanto, considerando a proporção de sementes germinadas no fim do período, as metodologias MLGM e EEG são as mais indicadas. Utilizou-se o MLGM2 para obter um gráfico das estimativas da probabilidade de germinação das sementes (Figura 19).

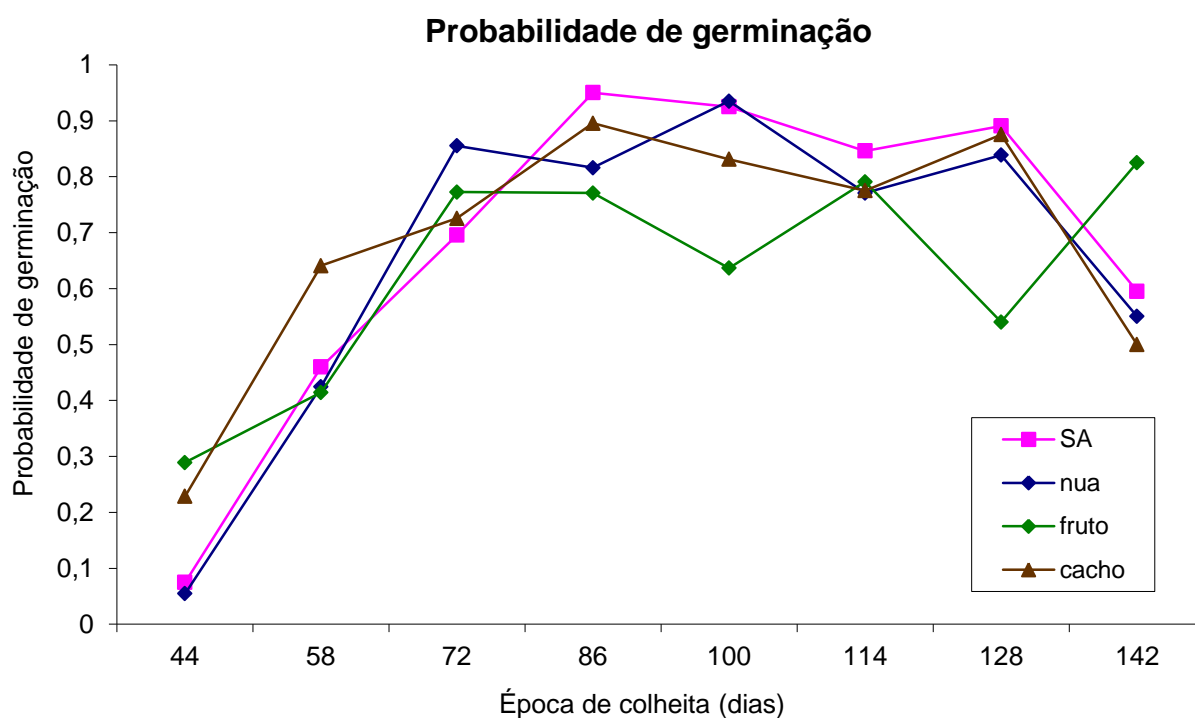


Figura 19: Gráfico das estimativas da probabilidade de germinação das sementes (MLGM2).

Na figura 19 observa-se que, considerando a proporção de sementes germinadas no final do período, a maior probabilidade de germinação ocorre no tratamento SA, na época 86. Em geral, para qualquer tratamento, colheitas antes dos 86 dias não produzem bons resultados. Há também uma tendência de queda na germinação em colheitas tardias, feitas após 114 dias.

### 6.3 Análise considerando a proporção de sementes germinadas por período

Considerando a proporção de sementes germinadas em cada período (7, 14 e 21 dias) e utilizando o modelo linear generalizado, visto na seção 4.4, com função de ligação cll (MLG3) e logit (MLG4), obtiveram-se os resultados apresentados na tabela 12.

Tabela 12: Resultados do ajuste do MLG3 e MLG4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLG3	MLG4	MLG3	MLG4	MLG3	MLG4
$\mu$	-3,2464	-3,3742	0,2595	0,2853	< 0,0001	< 0,0001
$\alpha_1$	0,7330	0,8334	0,2766	0,2979	0,0080	0,0051
$\alpha_2$	2,6674	3,1236	0,2505	0,2748	< 0,0001	< 0,0001
$\beta_1$	0,7025	0,8806	0,2050	0,2513	0,0006	0,0005
$\beta_2$	0,7724	0,9969	0,2024	0,2464	0,0001	< 0,0001
$\beta_3$	0,5958	0,8187	0,2162	0,2601	0,0058	0,0016
$\tau_1$	-0,7847	-0,6743	0,4461	0,4623	0,0786	0,1447
$\tau_2$	1,9898	2,2585	0,2734	0,3022	< 0,0001	< 0,0001
$\tau_3$	2,2242	2,5512	0,2686	0,3047	< 0,0001	< 0,0001
$\tau_4$	1,6861	1,8497	0,3132	0,3611	< 0,0001	< 0,0001
$\tau_5$	1,8554	2,2292	0,2963	0,3380	< 0,0001	< 0,0001
$\tau_6$	1,6818	1,9286	0,2944	0,3310	< 0,0001	< 0,0001
$\tau_7$	1,6334	1,7070	0,3027	0,3437	< 0,0001	< 0,0001
$(\alpha\beta)_{11}$	-0,7404	-0,9176	0,1812	0,2156	< 0,0001	< 0,0001
$(\alpha\beta)_{12}$	-0,9398	-1,1440	0,1778	0,2088	< 0,0001	< 0,0001
$(\alpha\beta)_{13}$	0,2504	0,2568	0,1964	0,2201	0,2023	0,2432
$(\alpha\beta)_{21}$	-0,3949	-0,4141	0,1614	0,1968	0,0144	0,0354
$(\alpha\beta)_{22}$	-0,5834	-0,7419	0,1561	0,1921	0,0002	0,0001
$(\alpha\beta)_{23}$	0,3178	0,2698	0,1781	0,2055	0,0744	0,1893
$(\alpha\tau)_{11}$	0,4700	0,4092	0,4734	0,4859	0,3208	0,3997

Tabela 12: Resultados do ajuste do MLG3 e MLG4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLG3	MLG4	MLG3	MLG4	MLG3	MLG4
$(\alpha\tau)_{21}$	-0,3074	-0,6422	0,4354	0,4481	0,4802	0,1518
$(\alpha\tau)_{12}$	-2,0976	-2,2293	0,3239	0,3418	< 0,0001	< 0,0001
$(\alpha\tau)_{22}$	-1,8368	-2,1281	0,2563	0,2779	< 0,0001	< 0,0001
$(\alpha\tau)_{13}$	-2,4332	-2,7170	0,3048	0,3284	< 0,0001	< 0,0001
$(\alpha\tau)_{23}$	-1,8067	-2,0057	0,2504	0,2813	< 0,0001	< 0,0001
$(\alpha\tau)_{14}$	-0,8225	-0,9627	0,3259	0,3586	0,0116	0,0073
$(\alpha\tau)_{24}$	-0,5864	-0,1969	0,2974	0,3403	0,0486	0,5629
$(\alpha\tau)_{15}$	-0,8086	-0,9627	0,3087	0,3414	0,0088	0,0048
$(\alpha\tau)_{25}$	-1,0246	-0,9423	0,2824	0,3245	0,0003	0,0037
$(\alpha\tau)_{16}$	0,5080	0,6650	0,2984	0,3237	0,0887	0,0399
$(\alpha\tau)_{26}$	-1,1959	-1,2613	0,2804	0,3122	< 0,0001	< 0,0001
$(\alpha\tau)_{17}$	-0,0370	0,0334	0,3092	0,3375	0,9048	0,9212
$(\alpha\tau)_{27}$	-0,8334	-0,6651	0,2864	0,3232	0,0036	0,0396
$(\beta\tau)_{11}$	-1,5110	-1,7010	0,3285	0,3574	< 0,0001	< 0,0001
$(\beta\tau)_{21}$	-1,7245	-1,8650	0,3649	0,3897	< 0,0001	< 0,0001
$(\beta\tau)_{31}$	-0,6154	-0,7901	0,2377	0,2778	0,0096	0,0045
$(\beta\tau)_{12}$	-0,9062	-1,1274	0,1981	0,2477	< 0,0001	< 0,0001
$(\beta\tau)_{22}$	-0,9763	-1,2086	0,2026	0,2489	< 0,0001	< 0,0001
$(\beta\tau)_{32}$	-1,4566	-1,7481	0,1958	0,2480	< 0,0001	< 0,0001
$(\beta\tau)_{13}$	-0,5289	-0,6924	0,1853	0,2449	0,0043	0,0047
$(\beta\tau)_{23}$	0,1103	0,1611	0,1837	0,2465	0,5482	0,5135
$(\beta\tau)_{33}$	-0,6897	-0,8417	0,1791	0,2243	0,0001	0,0006
$(\beta\tau)_{14}$	0,1312	0,4942	0,1826	0,2701	0,4724	0,0673
$(\beta\tau)_{24}$	-0,3866	-0,4723	0,1832	0,2605	0,0348	0,0698

Tabela 12: Resultados do ajuste do MLG3 e MLG4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLG3	MLG4	MLG3	MLG4	MLG3	MLG4
$(\beta\tau)_{34}$	-1,2877	-1,6310	0,1769	0,2558	< 0,0001	< 0,0001
$(\beta\tau)_{15}$	0,0597	-0,0682	0,1805	0,2527	0,7408	0,7874
$(\beta\tau)_{75}$	0,1560	0,0941	0,1816	0,2509	0,3903	0,7076
$(\beta\tau)_{35}$	-1,4127	-1,8345	0,1792	0,2464	< 0,0001	< 0,0001
$(\beta\tau)_{16}$	-0,3847	-0,6310	0,1847	0,2454	0,0373	0,0101
$(\beta\tau)_{76}$	-0,4156	-0,5792	0,1874	0,2453	0,0266	0,0182
$(\beta\tau)_{36}$	-1,0210	-1,2760	0,1806	0,2439	< 0,0001	< 0,0001
$(\beta\tau)_{17}$	0,0237	0,2573	0,1803	0,2502	0,8954	0,3038
$(\beta\tau)_{77}$	-0,3221	-0,3260	0,1822	0,2424	0,0770	0,1786
$(\beta\tau)_{37}$	-1,6847	-2,0297	0,1835	0,2429	< 0,0001	< 0,0001

Nas figuras 20 e 21 têm-se os gráficos dos resíduos de Pearson versus a ordem das observações para o MLG3 e MLG4, respectivamente.

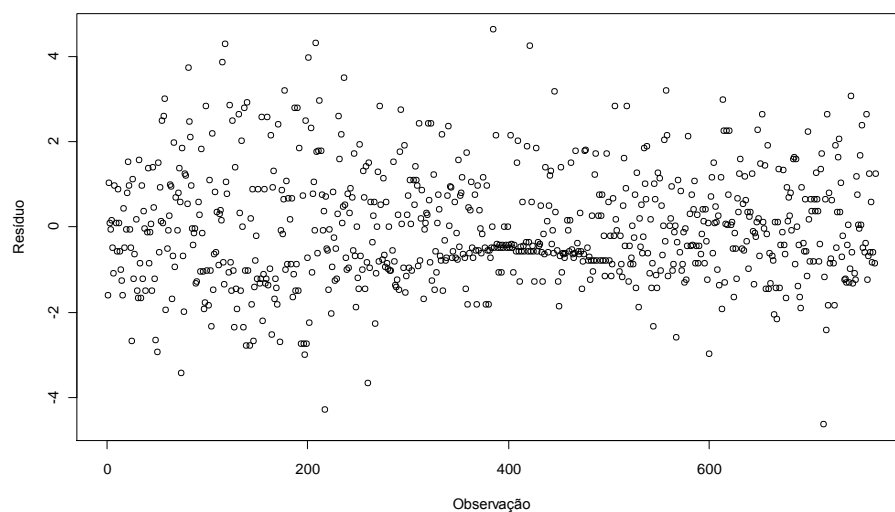


Figura 20: Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG3.

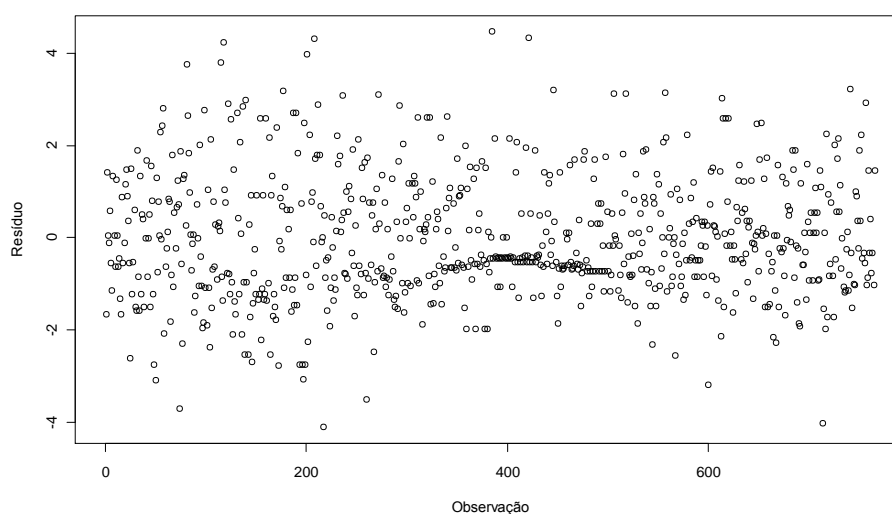


Figura 21: Gráfico dos resíduos de Pearson versus a ordem das observações para o MLG4.

Nas figuras 22 e 23 têm-se os gráficos de probabilidade normal para os resíduos de Pearson do MLG3 e MLG4, respectivamente.

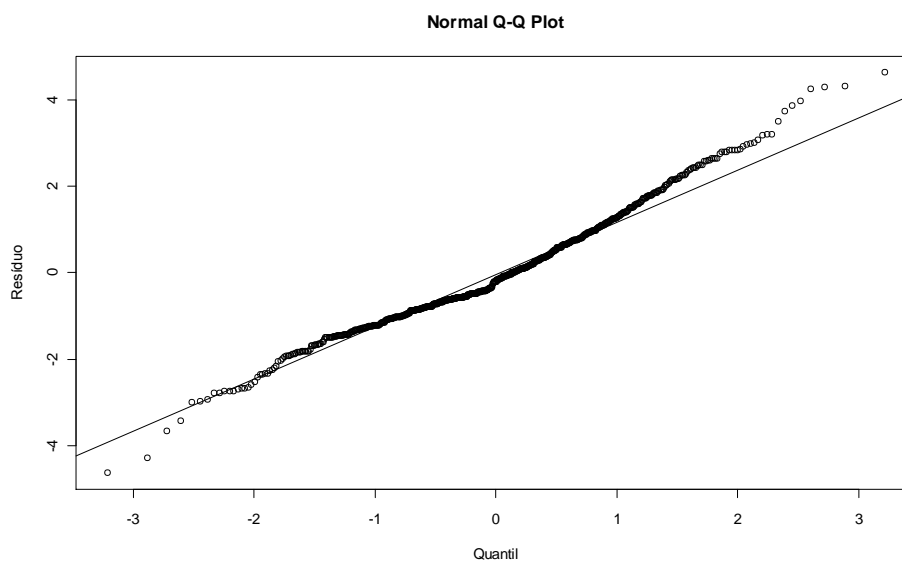


Figura 22: Gráfico de probabilidade normal para os resíduos de Pearson do MLG3.

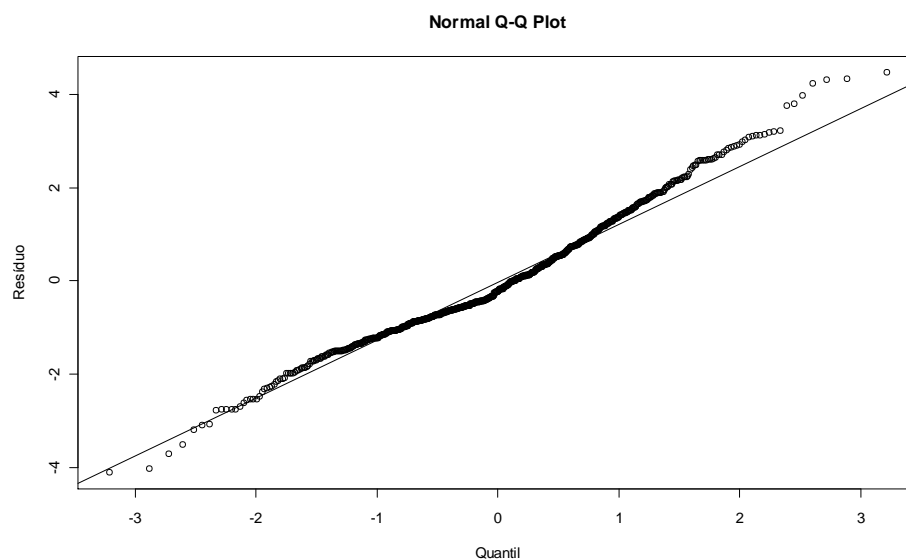


Figura 23: Gráfico de probabilidade normal para os resíduos de Pearson do MLG4.

Os gráficos dos resíduos de Pearson (Figuras 20 e 21) e os gráficos de probabilidade normal (Figuras 22 e 23) foram próximos, com os resíduos variando de -4 a 4,5.

Pelos gráficos de probabilidade normal (Figuras 22 e 23), verifica-se que os modelos ajustados (MLG3 e MLG4) não estão adequados. Uma solução provável seria a inclusão do efeito aleatório de rolo de papel no modelo.

Na tabela 13, têm-se os critérios de verificação do ajuste para o MLG3 com função de ligação complemento log-log e para o MLG4 com função de ligação logística.

Tabela 13: Critérios para verificação do ajuste do MLG3 e MLG4

Modelo	Critério	G.L.	Valor	Valor / G.L.
MLG3	<i>Deviance</i>	711	1341,6077	1,8869
	Qui-Quadrado de Pearson	711	1283,7266	1,8055
	AIC	—	3137,9169	—
MLG4	<i>Deviance</i>	711	1369,6228	1,9263
	Qui-Quadrado de Pearson	711	1318,6499	1,8546
	AIC	—	3165,9319	—

Comparando os critérios de verificação do ajuste do modelo (Tabela 13), verificam-se valores próximos para as duas funções de ligação utilizadas, sendo um pouco menor para a função de ligação complemento log-log.

Em todas as análises que consideram somente os efeitos fixos e sem levar em conta a dependência das observações, verificam-se valores altos nos critérios *Deviance*, Qui-Quadrado de Pearson e AIC.

Considerando a proporção de sementes germinadas em cada período (7, 14 e 21 dias), utilizando o modelo linear generalizado misto, visto na seção 4.6 e as funções de ligação complemento log-log (MLGM3) e logística (MLGM4), obtiveram-se os resultados apresentados na tabela 14.

Tabela 14: Resultados do ajuste do MLGM3 e MLGM4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLGM3	MLGM4	MLGM3	MLGM4	MLGM3	MLGM4
$\mu$	-3,2386	-3,3299	0,2919	0,3338	< 0,0001	< 0,0001
$\alpha_1$	0,7688	0,8401	0,3060	0,3470	0,0122	0,0157
$\alpha_2$	2,6580	3,1041	0,2819	0,3276	< 0,0001	< 0,0001
$\beta_1$	0,6680	0,7752	0,2690	0,3337	0,0133	0,0205
$\beta_2$	0,7998	0,9435	0,2651	0,3304	0,0026	0,0044
$\beta_3$	0,4750	0,6686	0,2785	0,3494	0,0886	0,0561
$\tau_1$	-0,7509	-0,6481	0,4757	0,5133	0,1149	0,2072
$\tau_2$	1,9794	2,2024	0,3196	0,3756	< 0,0001	< 0,0001
$\tau_3$	2,2420	2,5111	0,3166	0,3764	< 0,0001	< 0,0001
$\tau_4$	1,6918	1,8684	0,3586	0,4286	< 0,0001	< 0,0001
$\tau_5$	1,9442	2,1704	0,3388	0,4020	< 0,0001	< 0,0001
$\tau_6$	1,5067	1,6044	0,3453	0,4087	< 0,0001	< 0,0001
$\tau_7$	1,4473	1,5447	0,3559	0,4279	< 0,0001	0,0003
$(\alpha\beta)_{11}$	-0,5636	-0,6625	0,2243	0,2739	0,0122	0,0158



Tabela 14: Resultados do ajuste do MLGM3 e MLGM4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLGM3	MLGM4	MLGM3	MLGM4	MLGM3	MLGM4
$(\alpha\beta)_{12}$	-0,9751	-1,1487	0,2185	0,2690	< 0,0001	< 0,0001
$(\alpha\beta)_{13}$	0,2455	0,2758	0,2305	0,2745	0,2871	0,3153
$(\alpha\beta)_{21}$	-0,2965	-0,2725	0,2037	0,2589	0,1459	0,2930
$(\alpha\beta)_{22}$	-0,5581	-0,6447	0,1956	0,2512	0,0044	0,0105
$(\alpha\beta)_{23}$	0,3238	0,3036	0,2121	0,2614	0,1273	0,2459
$(\alpha\tau)_{11}$	0,3656	0,2631	0,4998	0,5358	0,4647	0,6236
$(\alpha\tau)_{21}$	-0,3362	-0,7086	0,4613	0,4956	0,4664	0,1533
$(\alpha\tau)_{12}$	-2,1693	-2,3128	0,3566	0,4016	< 0,0001	< 0,0001
$(\alpha\tau)_{22}$	-1,8644	-2,1707	0,2957	0,3475	< 0,0001	< 0,0001
$(\alpha\tau)_{13}$	-2,4781	-2,7103	0,3388	0,3873	< 0,0001	< 0,0001
$(\alpha\tau)_{23}$	-1,8245	-1,9909	0,2916	0,3507	< 0,0001	< 0,0001
$(\alpha\tau)_{14}$	-0,9529	-1,0486	0,3596	0,4162	0,0082	0,0120
$(\alpha\tau)_{24}$	-0,6124	-0,2289	0,3342	0,4022	0,0673	0,5694
$(\alpha\tau)_{15}$	-0,8697	-0,9712	0,3443	0,4004	0,0118	0,0155
$(\alpha\tau)_{25}$	-1,0257	-0,8470	0,3220	0,3884	0,0015	0,0296
$(\alpha\tau)_{16}$	0,5018	0,7168	0,3372	0,3913	0,1372	0,0674
$(\alpha\tau)_{26}$	-1,1800	-1,2537	0,3231	0,3840	0,0003	0,0011
$(\alpha\tau)_{17}$	0,0055	0,1119	0,3497	0,4093	0,9874	0,7847
$(\alpha\tau)_{27}$	-0,6723	-0,4681	0,3309	0,3998	0,0425	0,2420
$(\beta\tau)_{11}$	-1,5697	-1,7064	0,3890	0,4542	< 0,0001	0,0002
$(\beta\tau)_{21}$	-1,7856	-1,9171	0,4208	0,4836	< 0,0001	< 0,0001
$(\beta\tau)_{31}$	-0,5204	-0,7019	0,3156	0,3950	0,0996	0,0760
$(\beta\tau)_{12}$	-0,8545	-0,9825	0,2824	0,3596	0,0026	0,0064
$(\beta\tau)_{22}$	-1,0358	-1,2334	0,2871	0,3670	0,0003	0,0008
$(\beta\tau)_{32}$	-1,3102	-1,5805	0,2847	0,3698	< 0,0001	< 0,0001

Tabela 14: Resultados do ajuste do MLGM3 e MLGM4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	MLGM3	MLGM4	MLGM3	MLGM4	MLGM3	MLGM4
$(\beta\tau)_{13}$	-0,4838	-0,5777	0,2750	0,3563	0,0790	0,1054
$(\beta\tau)_{23}$	0,0062	0,0842	0,2752	0,3614	0,9821	0,8160
$(\beta\tau)_{33}$	-0,6163	-0,7859	0,2748	0,3664	0,0252	0,0323
$(\beta\tau)_{14}$	0,1745	0,3649	0,2747	0,3782	0,5253	0,3349
$(\beta\tau)_{24}$	-0,3781	-0,4766	0,2765	0,3707	0,1719	0,1990
$(\beta\tau)_{34}$	-1,1286	-1,4617	0,2751	0,3763	< 0,0001	0,0001
$(\beta\tau)_{15}$	-0,2201	-0,2430	0,2706	0,3647	0,4164	0,5056
$(\beta\tau)_{25}$	0,0025	0,0978	0,2705	0,3630	0,9927	0,7877
$(\beta\tau)_{35}$	-1,2829	-1,6418	0,2699	0,3609	< 0,0001	< 0,0001
$(\beta\tau)_{16}$	-0,3791	-0,4760	0,2737	0,3624	0,1665	0,1894
$(\beta\tau)_{26}$	-0,2506	-0,2947	0,2745	0,3590	0,3616	0,4120
$(\beta\tau)_{36}$	-0,7407	-0,8936	0,2728	0,3668	0,0068	0,0151
$(\beta\tau)_{17}$	0,0487	0,1140	0,2698	0,3653	0,8567	0,7550
$(\beta\tau)_{27}$	-0,1846	-0,2611	0,2719	0,3637	0,4973	0,4730
$(\beta\tau)_{37}$	-1,5135	-1,9420	0,2740	0,3689	< 0,0001	< 0,0001

Nas figuras 24 e 25 têm-se os gráficos dos resíduos de Pearson condicional versus a ordem das observações para o MLGM3 e MLGM4, respectivamente.

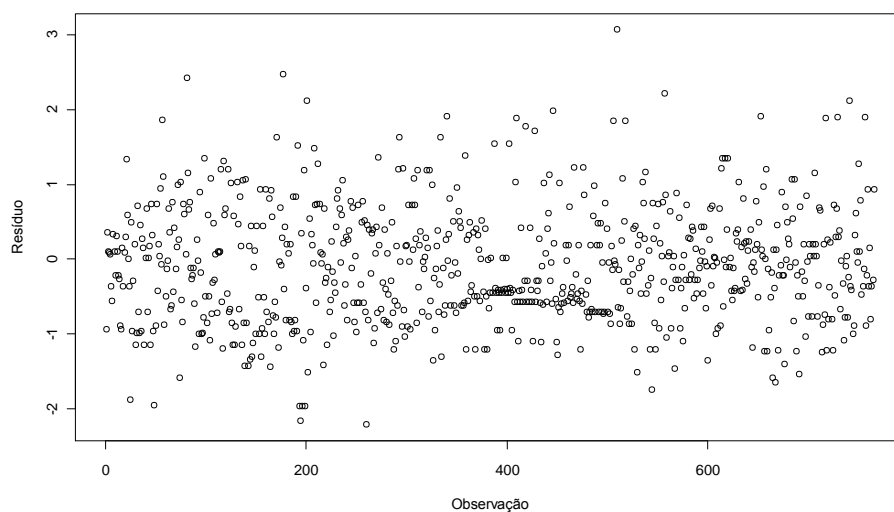


Figura 24: Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM3.

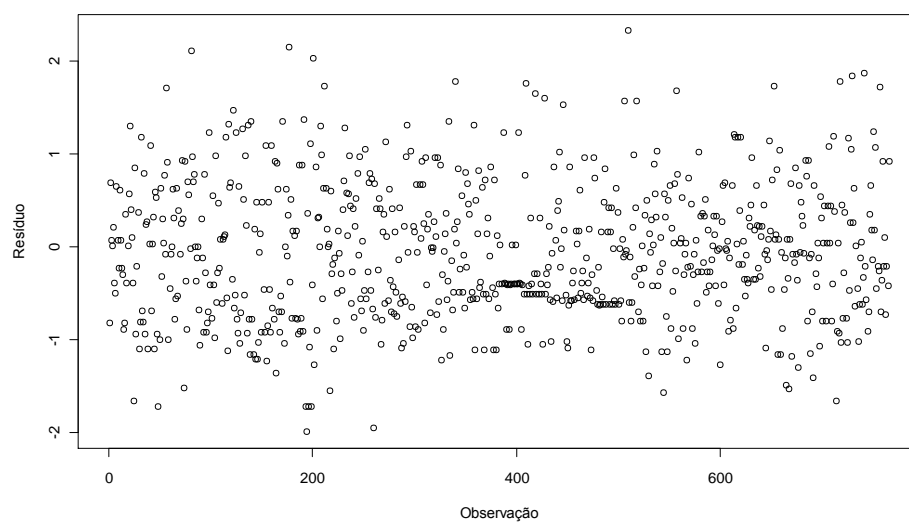


Figura 25: Gráfico dos resíduos de Pearson condicional versus a ordem das observações do MLGM4.

Nas figuras 26 e 27 têm-se os gráficos de probabilidade normal para os resíduos de Pearson condicional do MLGM3 e MLGM4, respectivamente.

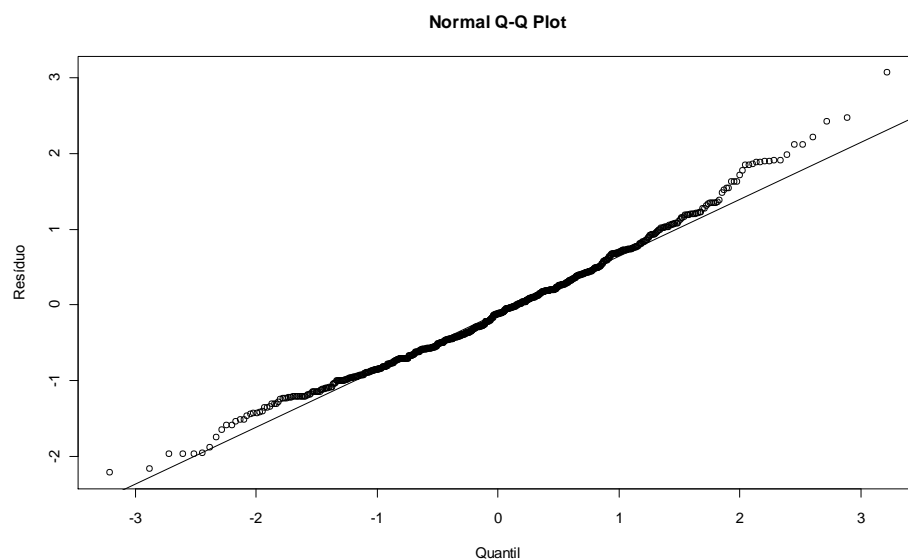


Figura 26: Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM3.

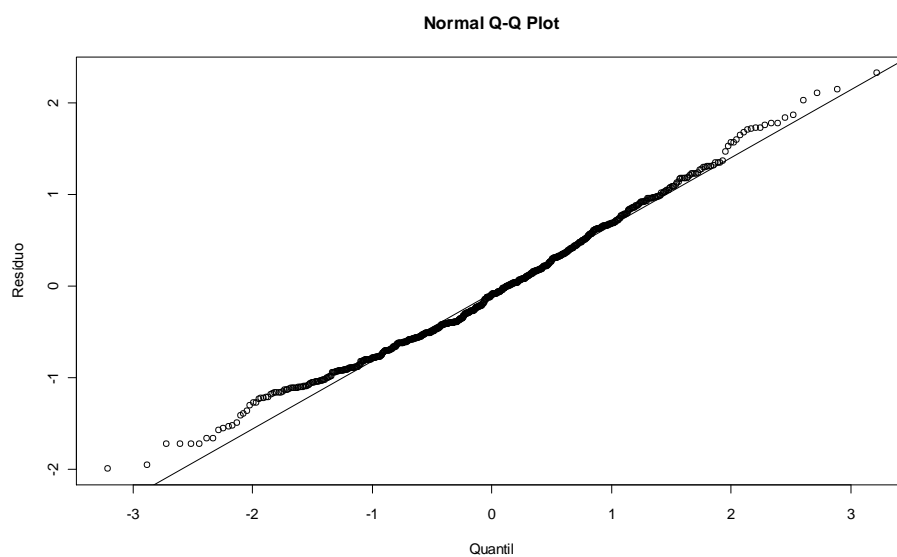


Figura 27: Gráfico de probabilidade normal para os resíduos de Pearson condicional do MLGM4.

Os gráficos dos resíduos de Pearson condicional (Figuras 24 e 25) e os gráficos de probabilidade normal (Figuras 26 e 27) foram próximos, com variação menor para o modelo com função de ligação logística.

Os critérios de verificação do ajuste do modelo (Tabela 15) apresentaram valores próximos, sendo um pouco menor para a função de ligação logística.

Tabela 15: Critério para verificação do ajuste do MLGM3 e MLGM4

Modelo	Critério	G.L.	Valor	Valor / G.L.
MLGM3	Qui-Quadrado Generalizado	711	736,18	1,0354
MLGM4	Qui-Quadrado Generalizado	711	701,50	0,9866

Utilizando a metodologia da equação de estimação generalizada com matriz de correlação não-estruturada, vista na seção 4.5, e as funções de ligação complemento log-log (EEG3) e logística (EEG4), obtiveram-se os resultados apresentados na tabela 16.

Tabela 16: Resultados do ajuste do EEG3 e EEG4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	EEG3	EEG4	EEG3	EEG4	EEG3	EEG4
$\mu$	-3,2464	-3,3742	0,2851	0,3265	< 0,0001	< 0,0001
$\alpha_1$	0,7330	0,8334	0,3206	0,3569	0,0225	0,0198
$\alpha_2$	2,6674	3,1236	0,2773	0,3142	< 0,0001	< 0,0001
$\beta_1$	0,7025	0,8806	0,2212	0,2824	0,0016	0,0019
$\beta_2$	0,7724	0,9969	0,2213	0,2811	0,0005	0,0004
$\beta_3$	0,5958	0,8187	0,2467	0,3071	0,0160	0,0079
$\tau_1$	-0,7847	-0,6743	0,4915	0,5141	0,1108	0,1901
$\tau_2$	1,9898	2,2585	0,2956	0,3369	< 0,0001	< 0,0001
$\tau_3$	2,2242	2,5512	0,2858	0,3304	< 0,0001	< 0,0001
$\tau_4$	1,6861	1,8497	0,3864	0,4612	< 0,0001	< 0,0001

Tabela 16: Resultados do ajuste do EEG3 e EEG4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	EEG3	EEG4	EEG3	EEG4	EEG3	EEG4
$\tau_5$	1,8554	2,2292	0,3408	0,3982	< 0,0001	< 0,0001
$\tau_6$	1,6818	1,9286	0,3649	0,4172	< 0,0001	< 0,0001
$\tau_7$	1,6334	1,7070	0,3733	0,4388	< 0,0001	0,0001
$(\alpha\beta)_{11}$	-0,7404	-0,9176	0,2347	0,2820	0,0017	0,0012
$(\alpha\beta)_{12}$	-0,9398	-1,1440	0,2344	0,2732	< 0,0001	< 0,0001
$(\alpha\beta)_{13}$	0,2504	0,2568	0,2520	0,2800	0,3207	0,3592
$(\alpha\beta)_{21}$	-0,3949	-0,4141	0,1884	0,2332	0,0364	0,0762
$(\alpha\beta)_{22}$	-0,5834	-0,7419	0,1847	0,2346	0,0016	0,0016
$(\alpha\beta)_{23}$	0,3178	0,2698	0,2268	0,2601	0,1615	0,3000
$(\alpha\tau)_{11}$	0,4700	0,4092	0,5437	0,5631	0,3877	0,4676
$(\alpha\tau)_{21}$	-0,3074	-0,6422	0,4951	0,5150	0,5350	0,2128
$(\alpha\tau)_{12}$	-2,0976	-2,2293	0,4009	0,4353	< 0,0001	< 0,0001
$(\alpha\tau)_{22}$	-1,8368	-2,1281	0,2717	0,2976	< 0,0001	< 0,0001
$(\alpha\tau)_{13}$	-2,4332	-2,7170	0,3345	0,3776	< 0,0001	< 0,0001
$(\alpha\tau)_{23}$	-1,8067	-2,0057	0,2748	0,3170	< 0,0001	< 0,0001
$(\alpha\tau)_{14}$	-0,8225	-0,9627	0,4068	0,4634	0,0435	0,0381
$(\alpha\tau)_{24}$	-0,5864	-0,1969	0,3717	0,4360	0,1150	0,6517
$(\alpha\tau)_{15}$	-0,8086	-0,9627	0,3684	0,4150	0,0285	0,0206
$(\alpha\tau)_{25}$	-1,0246	-0,9423	0,3245	0,3851	0,0017	0,0146
$(\alpha\tau)_{16}$	0,5080	0,6650	0,3811	0,4243	0,1830	0,1174
$(\alpha\tau)_{26}$	-1,1959	-1,2613	0,3495	0,3978	0,0007	0,0016
$(\alpha\tau)_{17}$	-0,0370	0,0334	0,3972	0,4461	0,9259	0,9404
$(\alpha\tau)_{27}$	-0,8334	-0,6651	0,3643	0,4339	0,0225	0,1258
$(\beta\tau)_{11}$	-1,5110	-1,7010	0,3142	0,3501	< 0,0001	< 0,0001
$(\beta\tau)_{21}$	-1,7245	-1,8650	0,4045	0,4292	< 0,0001	< 0,0001

Tabela 16: Resultados do ajuste do EEG3 e EEG4

Parâmetro	Estimativa		Erro-padrão		Valor p	
	EEG3	EEG4	EEG3	EEG4	EEG3	EEG4
$(\beta\tau)_{31}$	-0,6154	-0,7901	0,2247	0,2860	0,0063	0,0059
$(\beta\tau)_{12}$	-0,9062	-1,1274	0,1948	0,2685	< 0,0001	< 0,0001
$(\beta\tau)_{22}$	-0,9763	-1,2086	0,1901	0,2461	< 0,0001	< 0,0001
$(\beta\tau)_{32}$	-1,4566	-1,7481	0,2051	0,2785	< 0,0001	< 0,0001
$(\beta\tau)_{13}$	-0,5289	-0,6924	0,1937	0,2747	0,0065	0,0120
$(\beta\tau)_{23}$	0,1103	0,1611	0,1719	0,2370	0,5212	0,4969
$(\beta\tau)_{33}$	-0,6897	-0,8417	0,1707	0,2634	< 0,0001	0,0015
$(\beta\tau)_{14}$	0,1312	0,4942	0,2196	0,3181	0,5499	0,1208
$(\beta\tau)_{24}$	-0,3866	-0,4723	0,1801	0,2792	0,0322	0,0912
$(\beta\tau)_{34}$	-1,2877	-1,6310	0,1683	0,2747	< 0,0001	< 0,0001
$(\beta\tau)_{15}$	0,0597	-0,0682	0,2013	0,3184	0,7669	0,8306
$(\beta\tau)_{25}$	0,1560	0,0941	0,1969	0,2961	0,4285	0,7507
$(\beta\tau)_{35}$	-1,4127	-1,8345	0,2188	0,3281	< 0,0001	< 0,0001
$(\beta\tau)_{16}$	-0,3847	-0,6310	0,2587	0,3518	0,1375	0,0733
$(\beta\tau)_{26}$	-0,4156	-0,5792	0,2515	0,3367	0,0988	0,0859
$(\beta\tau)_{36}$	-1,0210	-1,2760	0,2234	0,3239	< 0,0001	< 0,0001
$(\beta\tau)_{17}$	0,0237	0,2573	0,2269	0,3310	0,9168	0,4373
$(\beta\tau)_{27}$	-0,3221	-0,3260	0,2765	0,3874	0,2444	0,4003
$(\beta\tau)_{37}$	-1,6847	-2,0297	0,1918	0,2920	< 0,0001	< 0,0001

Nas figuras 28 e 29 têm-se os gráficos dos resíduos de Pearson marginal versus a ordem das observações para o EEG3 e EEG4, respectivamente.

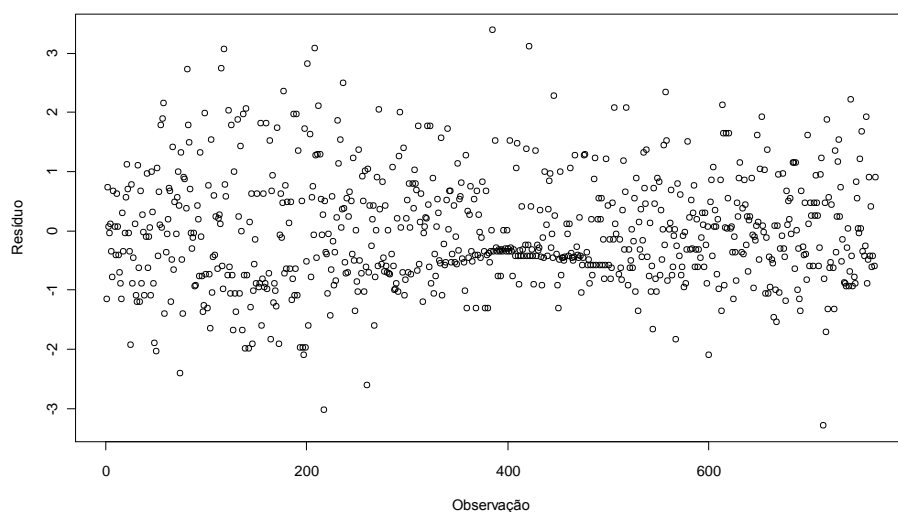


Figura 28: Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG3.

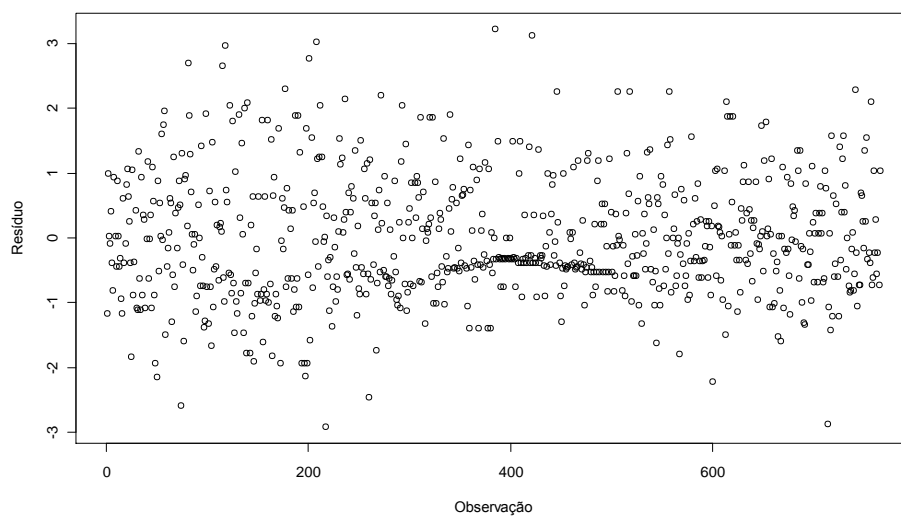


Figura 29: Gráfico dos resíduos de Pearson marginal versus a ordem das observações do EEG4

Nas figuras 30 e 31 têm-se os gráficos de probabilidade normal para os resíduos de Pearson marginal do EEG3 e EEG4, respectivamente.



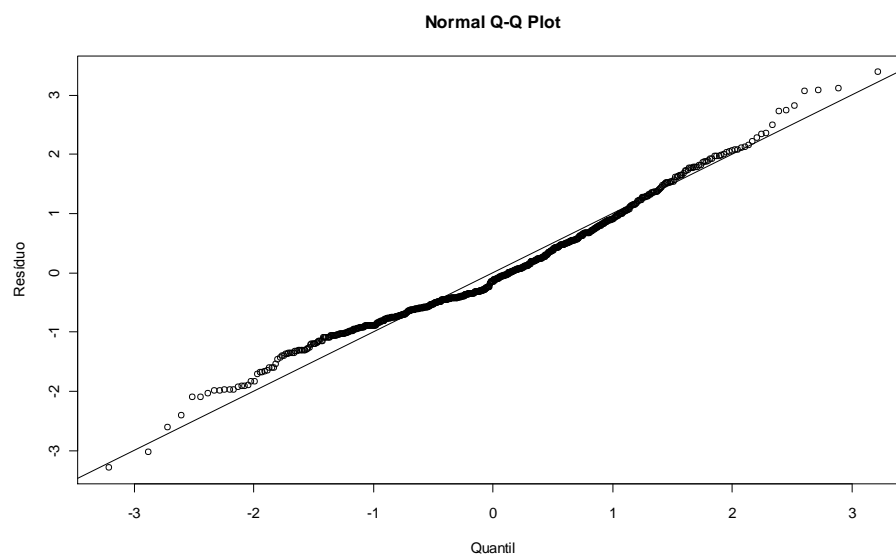


Figura 30: Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG3.

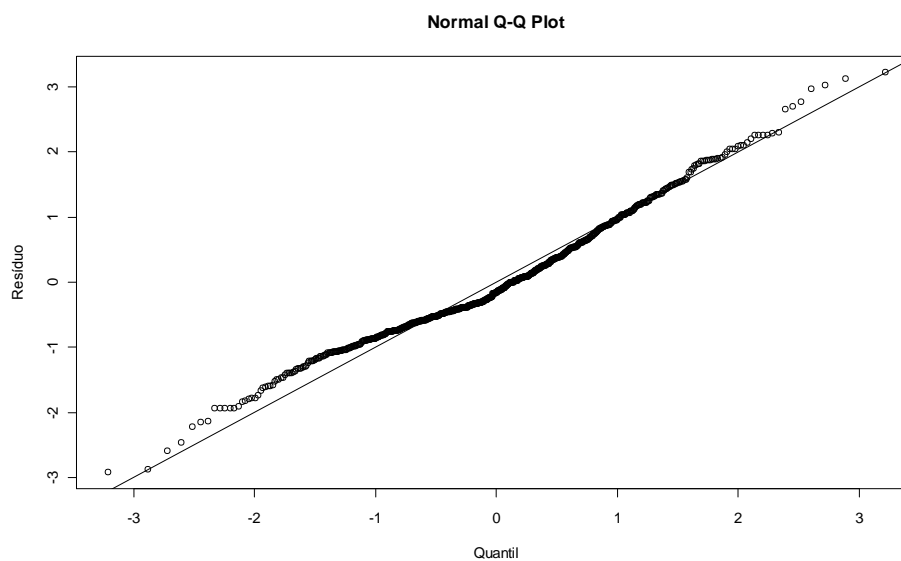


Figura 31: Gráfico de probabilidade normal para os resíduos de Pearson marginal do EEG4.

Os gráficos dos resíduos de Pearson marginal (Figuras 28 e 29) e os gráficos de probabilidade normal (Figuras 30 e 31) foram próximos. Observa-se uma variação menor para o modelo com função de ligação logística.

Os critérios de verificação do ajuste do modelo (Tabela 17) deram valores iguais a 1 para as duas funções de ligação utilizadas, indicando que os modelos se ajustaram adequadamente aos dados.

Tabela 17: Critério para verificação do ajuste EEG3 e EEG4

Modelo	Critério	G.L.	Valor	Valor / G.L.
EEG3	Qui-Quadrado Generalizado	711	711,00	1,0000
EEG4	Qui-Quadrado Generalizado	711	711,00	1,0000

Nas figuras 32, 33 e 34, têm-se as estimativas das probabilidades de germinação considerando o modelo linear generalizado misto com função de ligação logística (MLGM4). Observa-se que, considerando a proporção de sementes germinadas em cada um dos três períodos, a maior probabilidade de germinação ocorre no período 14, na época 86 do tratamento SA.

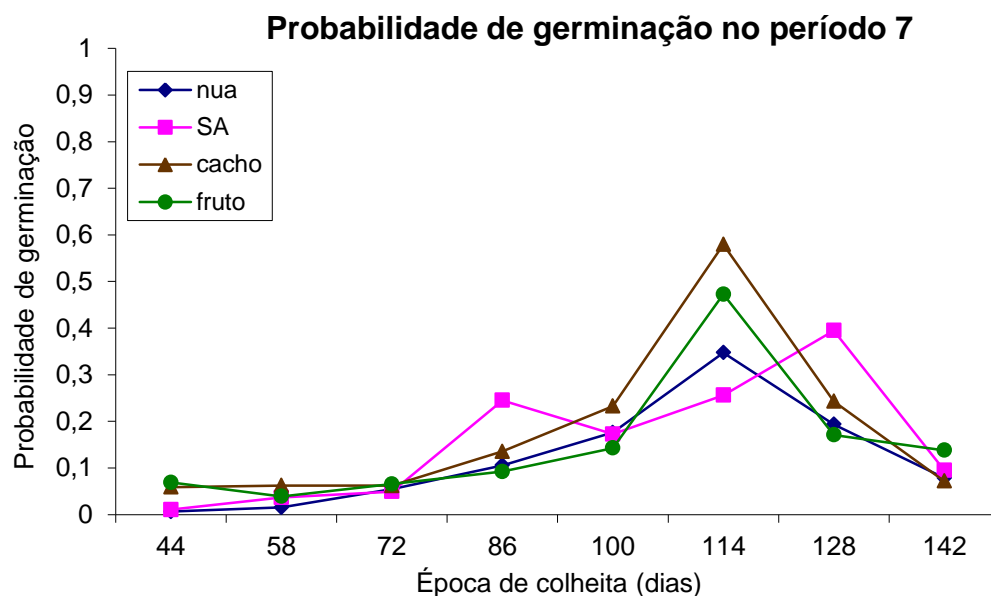


Figura 32: Gráfico das estimativas da probabilidade de germinação das sementes no período 7 (MLGM4).

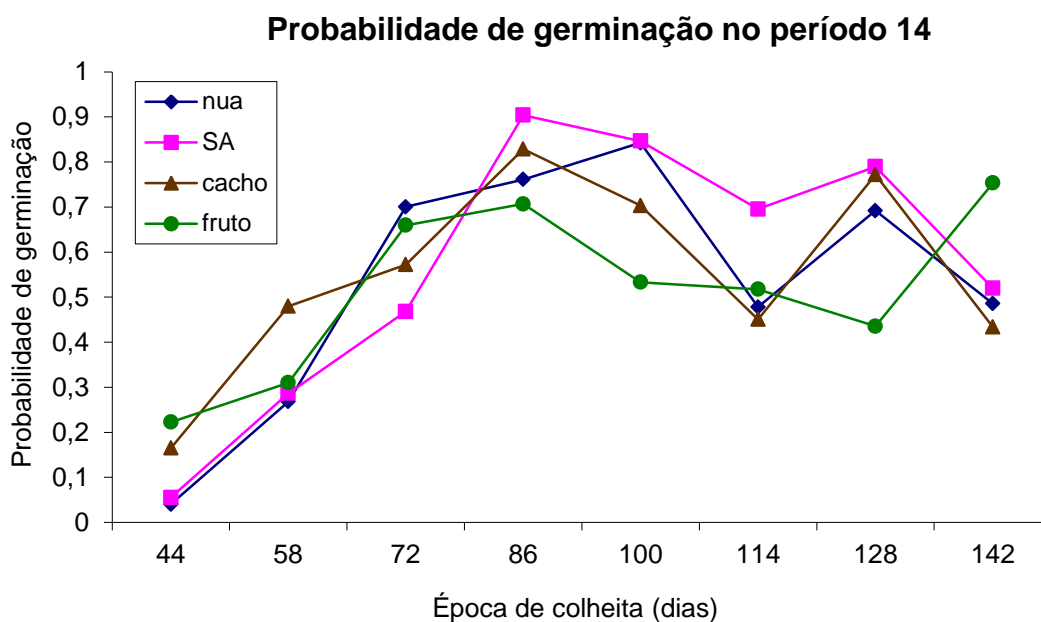


Figura 33: Gráfico das estimativas da probabilidade de germinação das sementes no período 14 (MLGM4).

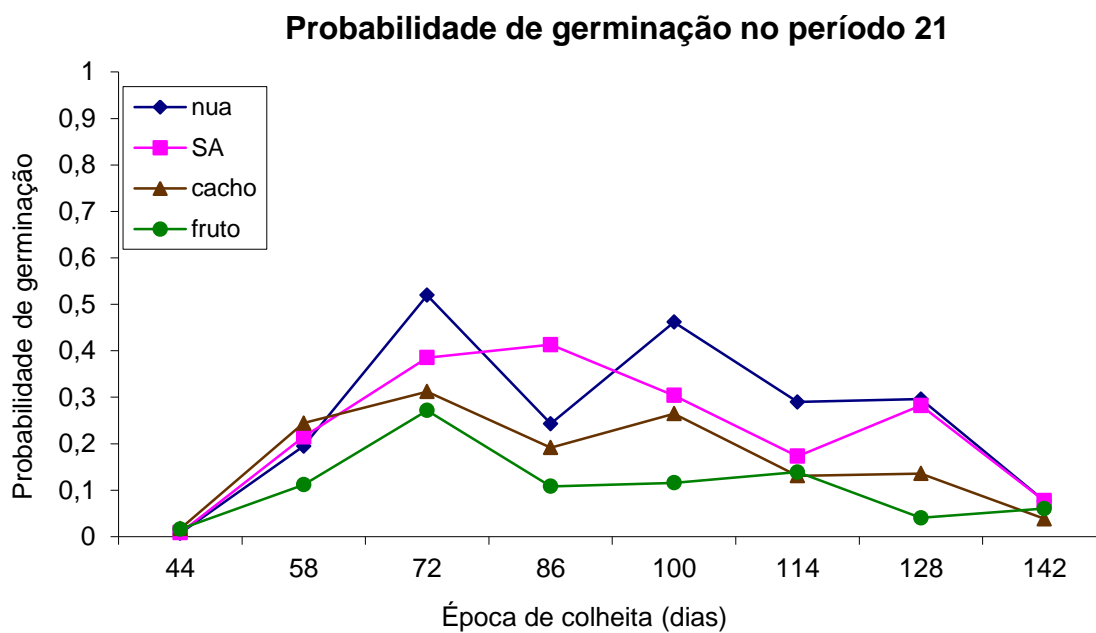


Figura 34: Gráfico das estimativas da probabilidade de germinação das sementes no período 21 (MLGM4).

Nas figuras 35 e 36 têm-se as curvas de germinação para cada época, considerando o modelo linear generalizado misto com função de ligação logística (MLGM4).

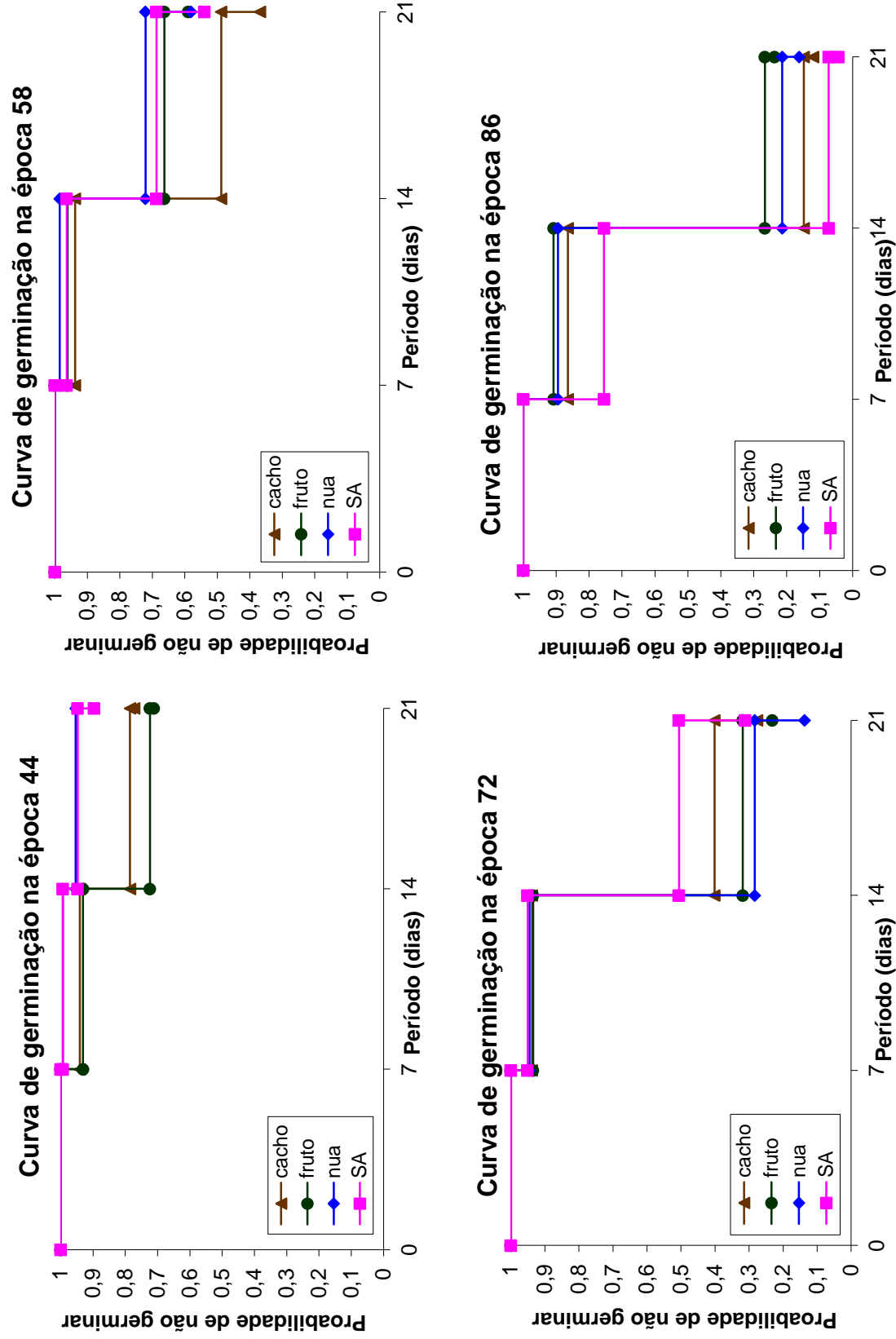


Figura 35: Curvas de germinação para as épocas 44, 58, 72 e 86 (MLGM4).

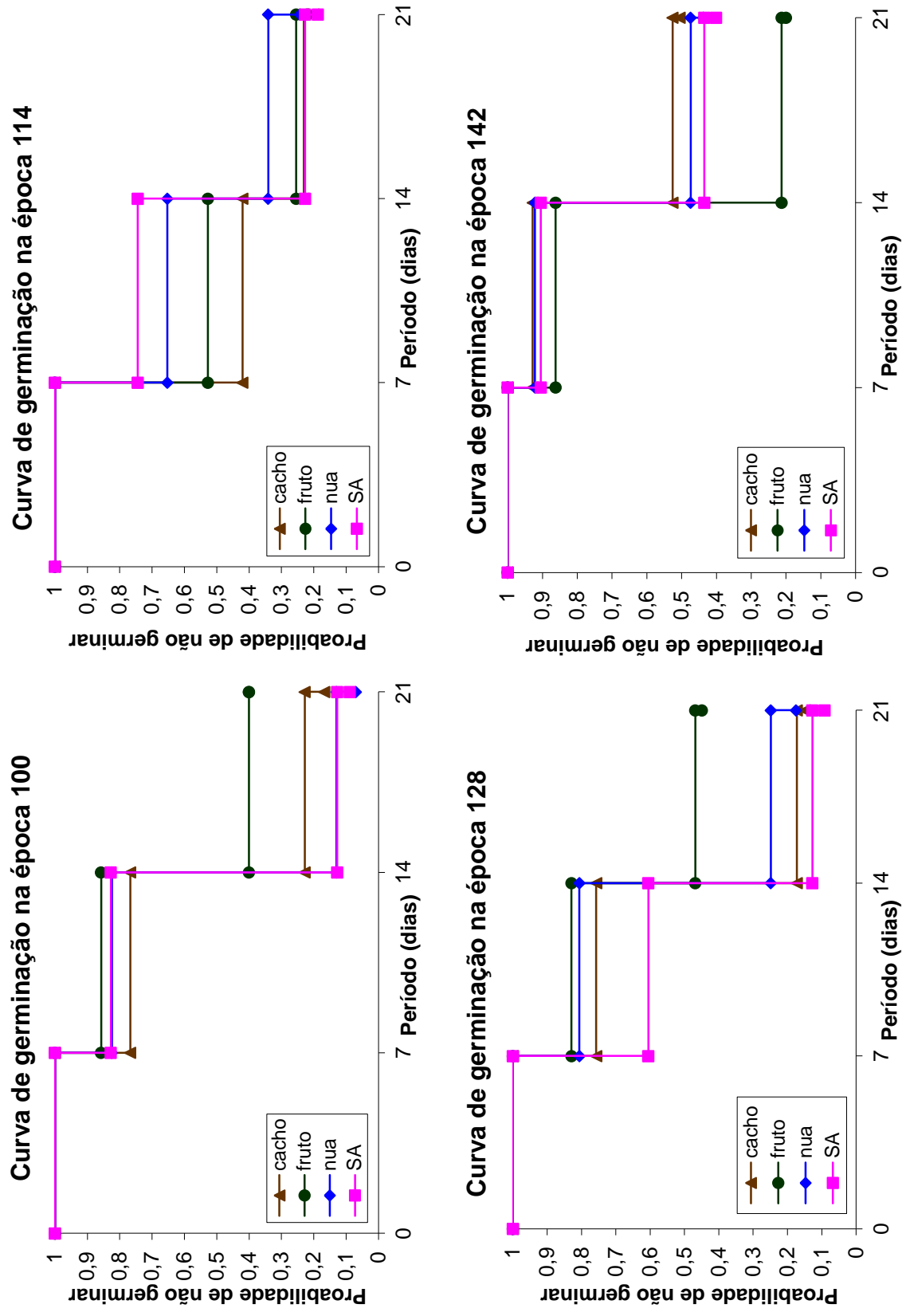


Figura 36: Curvas de germinação para as épocas 100, 114, 128 e 142 (MLGM4).

Nas figuras 35 e 36 observa-se que a probabilidade da semente não germinar é alta na época 44. Essa probabilidade vai diminuindo até a época 86 e começa a aumentar a partir da época 100. Na época 86, no período 14 do tratamento SA encontra-se a menor probabilidade da semente não germinar.

As sementes que não germinaram até os 14 dias, inclusive, dificilmente germinarão, principalmente se a época de colheita for tardia.

## 7 CONCLUSÕES

Os modelos lineares generalizados não se ajustaram adequadamente ao conjunto de dados (MGL1, MLG2, MLG3 e MLG4), pois eles consideram independência entre as observações, o que não se aplica aos dados do presente trabalho. Por esse motivo, as estatísticas *Deviance*, Qui-Quadrado de Pearson e AIC apresentam valores elevados, indicando que há uma superdispersão nos dados, ou seja, há uma variabilidade maior do que a explicada pela distribuição na família exponencial.

O ajuste dos modelos lineares generalizados mistos (MLGM1, MLGM2, MLGM3 e MLGM4) mostrou-se satisfatório, pois a inclusão do efeito aleatório de rolo de papel, colocado no modelo para captar a variabilidade existente nas observações correlacionadas, apresentou efeito, e a estatística Qui-Quadrado Generalizada resultou em valores próximos a 1 em todos eles.

No ajuste do método da equação de estimação generalizada (EEG1, EEG2, EEG3 e EEG4), o qual inclui uma matriz de correlação de trabalho no modelo, também se obteve bons resultados utilizando a matriz de correlação não-estruturada e as funções de ligação complemento log-log e logística. A estatística Qui-Quadrado Generalizada apresentou valores iguais a 1, indicando que o modelo está bem ajustado ao conjunto de dados.

Portanto, as metodologias MLGM e EEG são recomendadas quando se trabalha com dados correlacionados e estudos adicionais são necessários para a obtenção de mais critérios na seleção desses dois modelos.



## 8 REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, J. G. C. **Cultivar de mamona AL GUARANY 2002**. Boletim CATI, n.52, 2004.

ANDERSON, D. A.; AITKIN, M. Variance component models with binary response: Interviewer variability. **Journal of the Royal Statistical Society**, Series B, v.47, p.203–210, 1985.

AKAIKE, H. A new look at statistical model identification. **IEEE Transactions on Automatic Control**, v.19, p.716–722, 1974.

BRASIL. Ministério da Agricultura e Reforma Agrária. Secretaria Nacional de Defesa Agropecuária. Departamento Nacional de Defesa Vegetal. Coordenação de Laboratório Vegetal. **Regras para Análise de Sementes**. Brasília, DF, 1992. 365p.

CARNEIRO, J. W. P.; PIRES, J. C. Influência da temperatura e do substrato na germinação de sementes de mamona. **Revista Brasileira de Sementes**, v.5, n.3, p.127–131, 1983.

CARTAXO, W. V.; QUEIROGA, V. P.; FILHO, J. L. S.; QUEIROGA, D. A. N. Germinação de sementes de mamona com e sem casca. In: 3º Congresso Brasileiro de Mamona, 2008, Salvador.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de Sobrevida**. Rio de Janeiro: Editora FIOCRUZ, 2005, 396p.

CHALITA, L. V. A. S. **Modelos para dados agrupados e censurados**. 1997. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 1997. 135 f.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares generalizados**. Minicurso para o 12º SEAGRO e 52ª RBRAS. 165 p. Santa Maria: UFSM, 2007.

COSTA, S. C. **Modelos lineares generalizados mistos para dados longitudinais**. 2003. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2003. 110 f.

COX, D. R. Regression models and life tables (with discussion). **Journal of the Royal Statistical Society, Series B**, v.34, p.187–220, 1972.

COX, D. R.; SNELL, E. A general definition of residuals. **Journal of the Royal Statistical Society, Series B**, v.30, p.248–275, 1968.

DIGGLE, P. J.; HEAGERTY, P. J.; LIANG, K. Y.; ZEGER, S. L. **Analysis of Longitudinal Data**. New York: Oxford University Press, 2002, 379p.

DIGGLE, P. J.; LIANG, K. Y.; ZEGER, S. L. **Analysis of Longitudinal Data**. Oxford: Clarendon Press, 1994.

FAO. Food and Agriculture Organization of the United Nations. Disponível em: <<http://www.fao.org/>>. Acesso em: 18 nov. 2009.

FREIRE, E. C.; LIMA, E. F.; ANDRADE, F. P. Melhoramento genético. In: AZEVEDO, D. M. P.; LIMA, E. F. (Orgs.). **O agronegócio da mamona no Brasil**, Brasília, DF: EMBRAPA, Informação Tecnológica, 2001. p.229–256.

GEORGE, R. A. T. **Vegetable seed production**. 2.ed. London: CABI Publishing, 1985. 219p.

GREENWOOD, M. The natural duration of cancer. **Reports on Public Health and Medical Subjects**. London: Her Majesty's Stationery Office, v.33, p.1-26, 1926.

HORTON, N. J.; LIPSITZ, S. R. Review of Software to Fit Generalized Estimating Equation (GEE) Regression Models. **The American Statistician**, v.53, p.160–169, 1999.

HOLANDA, A. **Biodiesel e Inclusão Social**. Brasília: Câmara dos Deputados, 2004. 200p.

KALBFLEISCH, J. G. **Probability and Statistical Inference**. 2.ed. New York: Springer–Verlag, 1985.

KALBFLEISCH, J. G.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data**. New York: John Wiley & Sons, 1980. 321p.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observation. **Journal of the American Statistical Association**, v.53, p.457–481, 1958.

KIM, B. R.; CARTER, R.; RAO, P.; ARIET, M.; RESNICK, M. Standardized risk and description of results from multivariable modeling of a binary response. **Biometrical Journal**, v.48, n.1, p.54–66, 2006.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, v.38, p.963–974, 1982.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data**. New York: John Wiley & Sons, 1982. 580p.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear model (with discussion). **Journal of the Royal Statistical Society, Series B**, London, v.58, p.619–678, 1996.

LIANG, K. Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, v.73, n.1, p.13–22, 1986.

LINDSEY, J. K. **Applying Generalized Linear Models**. New York: Springer–Verlag, 1997.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D.; SCHABENBERGER, O. **SAS for Mixed Models**. 2.ed. Cary, USA: SAS Institute, 2006.

MALLARD, R. P. A mamona. **Boletim Agrícola**, Belo Horizonte, v.10, n. 3/12, p.111–117, 1961.

McCULLAGH, P. **Tensor Methods in Statistics**. New York: Chapman and Hall, 1987.

McCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2.ed. London: Chapman & Hall, 1989. 511p.

MENDES, R. C.; DIAS, D. C. F. S.; PEREIRA, M. D.; BERGER, P. G. Tratamentos pré-germinativos em sementes de mamona. **Revista Brasileira de Sementes**, v.31, n.1, p.187–194, 2009.

MOLENBERGHS, G.; VERBEKE, G. **Models for Discrete Longitudinal Data**. New York: Springer Science, 2005. 683p.

MOULTON, L. H.; ZEGER, S. L. Analyzing repeated measures on generalized linear models via the bootstrap. **Biometrics**, v.45, p.381–394, 1989.

NETER, J.; KUTNER, M. H.; NASCHSTHEIM, C. J.; WASSERMAN, W. **Applied Linear Statistical Models**. 4.ed. Chicago: McGraw Hill, 1996.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society, Series A**, London, v.135, p.370–384, 1972.

OCHI, Y.; PRENTICE, R. L. Likelihood inference in correlated probit regression. **Biometrika**, v.71, p.531–543, 1984.

PAN, W. Akaike's information criterion in generalized estimating equations. **Biometrics**, v.57, p.120–125, 2001.

PARENTE, E. J. S. **Biodiesel: uma aventura tecnológica num país engraçado**. Fortaleza, 2003. 66p.

PARK, T.; DAVIS, C. S.; LI, N. Alternative gee estimation procedures for discrete longitudinal data. **Computational Statistics & Data Analysis**, v.28, p.243–256, 1998.

PETROBRAS. Disponível em <<http://www.petrobras.com.br/pt/noticias/biodiesel-com-oleo-de-mamona-ja-e-realidade>>. Acesso em: 18 dez. 2009.

PREISSER, J. S.; QAQISH, B. F. Deletion diagnostics for generalized estimating equations. **Biometrika**, v.83, p.551–562, 1996.

PRENTICE, R. L.; GLOECKLER, L. A. Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. **Biometrics**, v.34, p.57–67, 1978.

PRENTICE, R. L.; ZHAO, L. P. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. **Biometrics**, v. 47, p.825–839, 1991.

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2010.

SANTOS, D. C.; CARVALHO, M. L. M.; OLIVEIRA, L. M.; KATAOKA, V. Y.; NETO, A. L. S. Teste de germinação em sementes de mamona. In: 2º Congresso Brasileiro de Mamona, 2006, Lavras.

SAS INSTITUTE. **SAS Software Release 9.2**. Cary, USA: SAS Institute, 2008.

SEAGRI Cultura Mamoneira. Disponível em: <<http://www.seagri.ba.gov.br/Mamoneira.htm>>. Acesso em: 14 dez. 2009.

SEBRAE. **Biodiesel**. Disponível em <[http://www.biblioteca.sebrae.com.br/bds/BDS.nsf/D170D324C7521915832572B200470F63/\\$File/NT00035116.pdf](http://www.biblioteca.sebrae.com.br/bds/BDS.nsf/D170D324C7521915832572B200470F63/$File/NT00035116.pdf)>. Acesso em: 15 mai. 2007.

SEVERINO, L. S.; GUIMARÃES, M. M. B.; COSTA, F. X.; LUCENA, A. M. A.; BELTRÃO, N. E. M.; CARDOSO, G. D. Emergência da plântula e germinação de semente de mamona plantada em diferentes posições. **Revista de Biologia e Ciências da Terra**, v.5, n.1, 2004.

SILVA, L. B. **Estágios de colheita e repouso pós-colheita dos frutos na qualidade de sementes de mamoneira**. 2007. Dissertação (Mestrado em Agronomia) – Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, 2007. 56 f.

SINGER, J. M.; ANDRADE, D. F. **Análise de Dados Longitudinais**. São Paulo: Campinas. VII Simpósio Nacional de Probabilidade e Estatística, 1986. 106p.

STRAM, D. O.; WEI, L. J.; WARE, J. H. Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. **Journal of the American Statistical Association**, v.83, n.403, p.631–637, 1988.

VENEZUELA, M. K. **Modelos lineares generalizados para análise de dados com medidas repetidas**. 2003. Dissertação (Mestrado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2003. 95 f.

VERBEKE, T.; DE CLERCQ, M. The Income-environment relationship: evidence from a binary response model. **Ecological Economics**, v.59, n.4, p.419–428, 2006.

VIEIRA, A. M. C. **Modelagem simultânea de média e dispersão e aplicações na pesquisa agrônoma**. 2008. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2008. 117 f.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. **Biometrika**, v.61, p.439–447, 1974.

WEIBULL, W. A statistical distribution of wide applicability. **Journal of Applied Mechanics**, v.18, p.293–297, 1951.

WHITEHEAD, J. The analysis of collapse clinical trials: with application to a comparison of two ulcer treatments. **Statistics in Medicine**, v.8, p.1439–1454, 1989.

ZEGER, S. L.; KARIM, M. R. Generalized linear model with random effects: gibbs sampling approach. **Journal of the American Statistical Association – Theory and Methods**, v.86, 1991.

ZEGER, S. L.; LIANG, K.Y. The analysis of discrete and continuous longitudinal data. **Biometrics**, v.42, p.121–130, 1986.

ZHAO, L. P.; PRENTICE, R. L. Correlated binary regression using a quadratic exponential model. **Biometrika**, v.77, p.642–648, 1990.

## APÊNDICE A

### O Método de Newton-Raphson

Suponhamos que desejamos encontrar a raiz  $\hat{\theta}$  da equação  $g(\theta) = 0$ . Seja  $\theta_0$  um valor fixo para  $\theta$ . Expandindo  $g(\theta)$ , através de uma série de Taylor no ponto  $\theta_0$ , temos

$$g(\theta) = g(\theta_0) + (\theta - \theta_0) g'(\theta_0) + \frac{(\theta - \theta_0)^2 g''(\theta_0)}{2!} + \dots$$

Para  $|\theta - \theta_0|$  pequeno, os  $n$ -ésimos termos na expressão acima são pequenos e, eliminando-se estes termos, temos

$$g(\theta) \approx g(\theta_0) + (\theta - \theta_0) g'(\theta_0).$$

Portanto, aproximamos  $g(\theta)$  por uma função linear de  $\theta$ . Como  $g(\theta) = 0$ , obtemos

$$g(\theta_0) + (\theta - \theta_0) g'(\theta_0) \approx 0,$$

ou,

$$\theta_1 \approx \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)}.$$

No método de Newton, tomamos  $\theta_0$  como sendo uma condição inicial para  $\theta$  e o valor de  $\theta$  na etapa seguinte,  $\theta_1$ , é dado pela expressão acima.



Repetindo os cálculos, tomando  $\theta_1$  como condição inicial, temos:

$$\theta_2 \approx \theta_1 - \frac{g(\theta_1)}{g'(\theta_1)}.$$

Continuamos este procedimento até  $\theta_{i+1} \approx \theta_i$ , neste caso,  $g(\theta_i) = 0$  e a raiz  $\hat{\theta}$  é dada por  $\theta_i$ .

No caso de dois parâmetros, precisamos resolver um par de equações simultâneas,

$$g(\alpha, \beta) = 0 \quad ; \quad h(\alpha, \beta) = 0.$$

Começando com um valor inicial  $(\alpha_0, \beta_0)$ , repetimos a expressão dada abaixo até obtermos a convergência, ou seja, até  $(\alpha_{i+1}, \beta_{i+1}) \approx (\alpha_i, \beta_i)$ .

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \approx \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} - \begin{bmatrix} \frac{\partial g}{\partial \alpha} & \frac{\partial g}{\partial \beta} \\ \frac{\partial h}{\partial \alpha} & \frac{\partial h}{\partial \beta} \end{bmatrix}^{-1} \begin{bmatrix} g \\ h \end{bmatrix}.$$

Esta generalização do método de Newton é chamada de método de Newton-Raphson (KALBFLEISCH, 1985).

## APÊNDICE B

### Comandos no SAS e no R

#### MLG1

```
proc genmod data=sasuser.luciano3;
class repet epoca cond;
model germi/n = epoca cond epoca*cond /err=bin link=cll;
output out=sasuser.cll p=y RESCHI=pearson RESDEV=dev STDRESCHI=pearpad
STDRESDEV=devpad;
run;
```

#### MLG2

```
proc genmod data=sasuser.luciano3;
class repet epoca cond;
model germi/n = epoca cond epoca*cond /err=bin link=logit ;
output out=sasuser.logit p=y RESCHI=pearson RESDEV=dev STDRESCHI=pearpad
STDRESDEV=devpad;
run;
```

#### MLG3

```
proc genmod data=sasuser.papel;
class repeticao epoca periodo condicao;
model germinadas1/n1 = periodo epoca condicao periodo*epoca periodo*condicao
epoca*condicao /err=bin link=cll ;
output out=sasuser.cll1 p=y RESCHI=pearson RESDEV=dev STDRESCHI=pearpad
STDRESDEV=devpad;
run;
```

**MLG4**

```
proc genmod data=sasuser.papel;
class repeticao epoca periodo condicao;
model germinadas1/n1 = periodo epoca condicao periodo*epoca periodo*condicao
epoca*condicao /err=bin link=logit;
output out=sasuser.logit1 p=y RESCHI=pearson RESDEV=dev STDRESCHI=pearpad
STDRESDEV=devpad; run;
```

**MLGM1**

```
proc glimmix data=sasuser.luciano3;
class repet epoca cond;
model germi/n = epoca cond epoca*cond /err=bin link=cll solution;
random repet*epoca*cond;
output out=sasuser.luc4 pred=predicted resid=r Pearson=pson STUDENT=st;
run;
```

**MLGM2**

```
proc glimmix data=sasuser.luciano3;
class repet epoca cond;
model germi/n = epoca cond epoca*cond /err=bin link=logit solution;
random repet*epoca*cond;
output out=sasuser.luc3 pred=predicted resid=r Pearson=pson STUDENT=st;
run;
```

**MLGM3**

```
proc glimmix data=sasuser.papel;
class repeticao periodo epoca condicao;
model germinadas1/n1 = periodo epoca condicao periodo*epoca periodo*condicao
epoca*condicao /err=bin link=cll solution;
random repeticao / subject=epoca*condicao*periodo;
output out=sasuser.luc2 pred=predicted resid=r Pearson=pson STUDENT=st; run;
```

**MLGM4**

```
proc glimmix data=sasuser.papel;
class repeticao periodo epoca condicao;
model germinadas1/n1 = periodo epoca condicao periodo*epoca periodo*condicao
epoca*condicao /err=bin link=logit solution;
random repeticao / subject=epoca*condicao*periodo;
output out=sasuser.luc1 pred=predicted resid=r Pearson=pson STUDENT=st; run;
```

**EEG1**

```
proc glimmix data=sasuser.luciano3;
class repet epoca cond;
model germi/n = epoca cond epoca*cond /err=bin link=cll solution;
random _residual_ /subject=repet*epoca*cond type=un vcorr;
output out=sasuser.geefincll pred=predicted resid=r Pearson=pson STUDENT=st;
run;
```

**EEG2**

```
proc glimmix data=sasuser.luciano3;
class repet epoca cond;
model germi/n = epoca cond epoca*cond /err=bin link=logit solution;
random _residual_ /subject = repet*epoca*cond type=un vcorr;
output out=sasuser.geefinlog pred=predicted resid=r Pearson=pson STUDENT=st;
run;
```

**EEG3**

```
proc glimmix data=sasuser.papel empirical;
class repeticao periodo epoca condicao;
model germinadas1/n1 = periodo epoca condicao periodo*epoca periodo*condicao
epoca*condicao /err=bin link=cll s;
random _residual_ / subject=repeticao*periodo*epoca*condicao type=un vcorr;
output out=sasuser.geecll pred=predicted resid=r Pearson=pson STUDENT=st; run;
```

**EEG4**

```

proc glimmix data=sasuser.papel empirical;
class repeticao periodo epoca condicao;
model germinadas1/n1 = periodo epoca condicao periodo*epoca periodo*condicao
epoca*condicao /err=bin link=logit s;
random _residual_ / subject=repeticao*periodo*epoca*condicao type=un vcorr;
output out=sasuser.geelogit pred=predicted resid=r Pearson=pson STUDENT=st;
run;

```

**Gráfico de Kaplan-Meier no R para cada tratamento**

```

dados<- read.table("Mamona.csv",h=T,sep=";")
require(survival)
par(mfrow=c(2,2))
resp <-Surv(dados$tempo[dados$tratamento=='SA'],dados$status[dados$tratamento=='SA'])
fit <- survfit(formula=resp~epoca, data=dados[dados$tratamento=='SA',])
plot(fit, conf.int=F, xlab="Tempo (dias)", ylab="S(T)",
mark.time=F,col=c(1:9),lty=c(rep(1,8),2),main='SA')
legend(0.1,0.8,c("30","44","58","72","86","100","114","128","142"),lty=c(rep(1,8),2),col=c(1:
9),cex=0.6,bty='n')
resp <-
Surv(dados$tempo[dados$tratamento=='cacho'],dados$status[dados$tratamento=='cacho'])
fit <- survfit(formula=resp~epoca, data=dados[dados$tratamento=='cacho',])
plot(fit, conf.int=F, xlab=" Tempo (dias)", ylab="S(T)",
mark.time=F,col=c(1:9),lty=c(rep(1,8),2),main='cacho')
legend(0.1,0.8,c("30","44","58","72","86","100","114","128","142"),lty=c(rep(1,8),2),col=c(1:
9),cex=0.6,bty='n')
resp <-Surv(dados$tempo[dados$tratamento=='nua'],dados$status[dados$tratamento=='nua'])
fit <- survfit(formula=resp~epoca, data=dados[dados$tratamento=='nua',])
plot(fit, conf.int=F, xlab=" Tempo (dias)", ylab="S(T)",
mark.time=F,col=c(1:9),lty=c(rep(1,8),2),main='nua')

```

```

legend(0.1,0.8,c("30","44","58","72","86","100","114","128","142"),lty=c(rep(1,8),2),col=c(1:
9),cex=0.6,bty='n')
resp <-
Surv(dados$tempo[dados$tratamento=='fruto'],dados$status[dados$tratamento=='fruto'])
fit <- survfit(formula=resp~epoca, data=dados[dados$tratamento=='fruto',])
plot(fit, conf.int=F, xlab=" Tempo (dias)", ylab="S(T)",
mark.time=F,col=c(1:9),lty=c(rep(1,8),2),main='fruto')
legend(0.1,0.8,c("30","44","58","72","86","100","114","128","142"),lty=c(rep(1,8),2),col=c(1:
9),cex=0.6,bty='n')

```