

UNIVERSIDADE DE BRASÍLIA  
INSTITUTO DE CIÊNCIAS EXATAS  
DEPARTAMENTO DE ESTATÍSTICA

Alan Renner Borges Nascimento

*Análise da produtividade de cultivares de café utilizando  
modelos lineares mistos*

Brasília  
2016

Alan Renner Borges Nascimento

*Análise da produtividade de cultivares de café utilizando  
modelos lineares mistos*

Monografia apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília, como requisito para a obtenção do grau de BACHAREL em Estatística.

**Orientador: Eduardo Monteiro de Castro Gomes**

**Doutor em Ciências Estatísticas - USP**

**Co-orientador: Juaci Vitoria Malaquias**

**Mestre em Estatística - UnB**

Brasília

2016

*Aos meus pais,  
Terezinha e Toninho,  
dedico.*

# Resumo

## **Análise da produtividade de cultivares de café utilizando modelos lineares mistos**

Este trabalho se propôs a abordar o uso de modelos lineares mistos para a análise de dados com medidas repetidas no tempo. O processo de modelagem iniciou com a escolha dos efeitos aleatórios em seguida dos efeitos fixos e da melhor estrutura de covariâncias. Para selecionar a estrutura da matriz de covariâncias que melhor se ajustou aos dados, utilizou-se os critérios de informação de Akaike - AIC e de Schwarz - BIC e o Teste da Razão de Verossimilhança. A análise gráfica dos resíduos foi utilizada para verificar o ajuste do modelo. Este trabalho utilizou dados de produtividade de um experimento de café Arábica realizado na Embrapa do Gama-DF, sendo que os resultados foram obtidos pelo programa estatístico R.

Palavras-chaves: Modelo linear Misto; Medidas repetidas; Café; Estrutura de Covariância.

# Abstract

## Coffee cultivars productivity analysis using mixed linear models

This work proposes to address the use of mixed linear models for data analysis with repeated measures. The modeling process began with the selection of random effects then the fixed effects and best covariance structure. To select the structure of the covariance matrix that best fit the data, we used the Akaike information criterion - AIC and Schwarz - BIC and the Likelihood Ratio Test. The graphical analysis of the waste was used to verify the model fit. This study used data from a productivity Arabica coffee experiment conducted at Embrapa Gama -DF and the results were obtained by the statistical program R.

Keywords : Mixed linear model; Repeated measures; Coffee; Covariance structure.

## Agradecimentos

À minha família que sempre me apoiou e esteve ao meu lado durante essa longa jornada.

À minha irmã Alaís Borges Nascimento que sempre esteve presente, dividindo a mesa de estudos e passando confiança para a conclusão deste trabalho.

À Universidade de Brasília, especialmente ao Departamento de estatística, pela oportunidade de realização do curso.

Ao Prof. Dr. Eduardo Monteiro de Castro Gomes, pela orientação, amizade e principalmente, pela paciência durante todo o trabalho.

Ao Prof. Dr. Lúcio José Vivaldi, pela orientação, pela atenção em momentos de dificuldade e pela paciência em passar seu conhecimento.

Ao Dr. Adriano Delly Veiga e Mestre Juaci Vitoria Malaquias que ajudaram a definir o objetivo deste trabalho e disponibilizaram o conjunto de dados para sua realização.

Aos professores do Departamento de Estatística pelos seus ensinamentos.

E a todos que, de forma direta ou indireta, contribuíram para a realização deste trabalho.

# Sumário

<b>Lista de Figuras</b>	<b>7</b>
<b>Lista de Tabelas</b>	<b>8</b>
<b>1 Introdução</b>	<b>9</b>
<b>2 Revisão de Literatura</b>	<b>11</b>
2.1 O café . . . . .	11
2.2 Medidas Repetidas . . . . .	12
2.3 Modelos Lineares Mistos . . . . .	14
2.4 Estruturas das Matrizes de covariância . . . . .	15
2.5 Estimação . . . . .	18
2.6 Seleção do modelo . . . . .	19
2.6.1 Critério de informação de Akaike . . . . .	20
2.6.2 Critério de informação bayesiano . . . . .	20
2.6.3 Teste de Wald . . . . .	21
2.6.4 Teste da Razão de verossimilhança . . . . .	22
<b>3 Material e Métodos</b>	<b>23</b>
3.1 Material . . . . .	23
3.2 Métodos . . . . .	27
<b>4 Resultados e Discussão</b>	<b>30</b>
<b>5 Considerações finais</b>	<b>40</b>

Referências Bibliográficas	41
Apêndices	
A Programa no Software R para o ajuste do modelo linear misto	45



## Lista de Figuras

3.1	Perfis individuais da produtividade do café, por cultivar. . . . .	26
3.2	Perfis médios da produtividade do café, por cultivar. . . . .	27
4.1	Boxplot da produtividade do café em sacas por hectare (sc/ha), por cultivar.	30
4.2	Boxplot da produtividade do café em sacas por hectare (sc/ha), por ano. .	31
4.3	Ajuste do modelo M1.4 aos dados de produtividade do café . . . . .	37
4.4	Gráfico de diagnóstico dos resíduos condicionais padronizados . . . . .	38
4.5	Gráficos de diagnóstico dos resíduos . . . . .	39

## Lista de Tabelas

3.1	Relação das cultivares de café com suas respectivas médias e desvios padrões (D.P) da produtividade . . . . .	24
3.2	Dados obtidos do experimento de café em função dos tratamentos e medidas realizadas ao longo do tempo para a variável produtividade. . . . .	25
4.1	Valores médios, desvios padrão, mínimos, máximos de produtividade do café nos quatro anos. . . . .	32
4.2	Correlações (abaixo da diagonal), variâncias (diagonal) e covariâncias (acima da diagonal) amostrais, da produtividade do café, por ano. . . . .	32
4.3	Estruturas dos efeitos fixos, aleatórios e seu método de estimação, para dados de produtividade do café. . . . .	33
4.4	Comparação dos modelos M1.1, M2.1 e M3.1 pelos critérios de informação AIC e BIC e pelo teste da razão verossimilhança (TRV) . . . . .	33
4.5	Teste de Wald para os parâmetros de efeito fixo do modelo M1.1 . . . . .	34
4.6	Comparação dos modelos M1.2, M1.3 pelo teste da razão verossimilhança (TRV) . . . . .	34
4.7	Comparação dos modelos M1.4, M1.5 e M1.6 pelos critérios AIC, BIC e teste da razão verossimilhança (TRV) . . . . .	35
4.8	Resultados do teste de Tukey para comparação de médias das cultivares com maior produtividade média . . . . .	36

# 1 Introdução

O Brasil é o maior produtor e exportador de café do mundo. Durante a safra de 2014 foram produzidas 45.341,62 milhões de sacas de 60 quilos em uma área de 1.947.199,80 hectares (CONAB, 2014), representando aproximadamente 32,16% da produção mundial e 6,9% das exportações do agronegócio brasileiro (MAPA, 2014).

A produtividade do café é afetada por diversos fatores dentre eles podemos citar a cultivar, condições climáticas e o manejo. Estudar a resposta dos diferentes cultivares em cada ambiente de produção nos auxilia a maximizar a exploração da cultura.

Esses fatores auxiliam a escolha da cultivar de café que melhor se adapta em cada região do país, dessa maneira é possível adequar o manejo e a cultivar para localidades específicas para obter melhores rendimentos e consequentemente maior lucratividade para os produtores rurais.

A Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) estuda as diversas cultivares de café em busca das mais produtivas e adaptadas para a realidade de cada região do país. O estudo realizado por Veiga et al. (2015) no experimento de café instalado em 2007 na Embrapa Hortaliças, Gama-DF, tinha como objetivo avaliar diversos genótipos de café arábica resistentes à ferrugem na região do cerrado do planalto central.

Os dados do experimento foram coletados entre os anos de 2010 a 2013 e foram analisados utilizando um modelo linear com efeitos fixos, neste contexto, uma abordagem mais apropriada a esse experimento seria considerar efeitos aleatórios além dos efeitos fixos no modelo.

O uso de modelos lineares mistos tem despertado grande interesse na estatística e em outras áreas, pois possibilitam a modelagem da correlação intra-indivíduo, não raramente encontrada em experimentos com medidas repetidas no tempo. Os dados da variável resposta do mesmo objeto observada repetidas vezes não podem ser consideradas não correlacionados, sendo assim uma boa maneira de analisar esse conjunto de dados com dependência intra-indivíduos seria os modelos lineares mistos (PINHEIRO, 1994).

Os modelos lineares mistos apresentam efeitos fixos e efeitos aleatórios, permitindo descrever a tendência temporal, apresentar estimadores mais precisos e verificar

---

alterações em decorrência do tempo. (FAUSTO et al.,2008).

Como última observação, a estrutura desse trabalho está dividida da seguinte maneira: Foi realizada a revisão de literatura em história do café, análise de dados longitudinais e modelos lineares mistos, na sessão metodologia o experimento é detalhado e por fim os resultados obtidos na análise utilizando modelos lineares mistos.

## 2 Revisão de Literatura

### 2.1 O café

Não existem evidências reais a respeito da descoberta do café, porém algumas lendas relatam sua possível origem. Uma das lendas mais aceitas é a de um criador de cabras na Etiópia, cerca de mil anos atrás ao observar seus animais, percebeu que após a ingestão dos frutos amarelo-avermelhado ficavam saltitantes e alegres. Essa planta é originária da Etiópia, onde, ainda nos tempos atuais faz parte na vegetação natural. O grande responsável pela propagação do café foi a Arábia. Durante o século XIV o café era conhecido como “vinho da Arábia”(ABIC, 2011).

Segundo a Associação Brasileira da Indústria de Café (ABIC, 2011), durante muitos anos o café foi utilizado como estimulante para rebanhos ao longo de viagens, sendo consumida em sua forma natural e fresca. Foi em 1000 d.C., que os árabes iniciaram o preparo da bebida café com a fervura das cerejas, porém apenas no século XIV que o processo de torrefação foi aplicado deixando assim a bebida com aspecto próximo ao que temos hoje.

O café pertence à família Rubiaceae, gênero *Coffea* L. e existem de 80 a 100 espécies. Apesar da grande quantidade, no Brasil temos apenas duas importantes espécies exploradas comercialmente que são: *Coffea arábica* e *Coffea canéfora* (ARTERO, 2006).

Uma importante característica presente na produção do cafeeiro é a bienalidade. Esse fenômeno produz a alternância de altas e baixas produções ao longo do tempo. Pode-se explicar esse efeito devido ao desgaste das plantas em anos com altas produções, trazendo como consequência a redução no ano seguinte. De acordo com Mendonça et al.(2011), o efeito da bienalidade pode estar relacionado com a constituição genética das plantas, ou seja, alguns genótipos podem apresentar diferentes características produtivas, por exemplo: alta variabilidade ou estabilidade na produção anual de grãos.

A planta foi trazida para o Brasil em 1727, primeiramente foi difundida no litoral brasileiro e a partir daí devido às boas condições climáticas se expandiu em todas as regiões do país. Desde o século XIX a produção de café contribui significativamente

para economia.

Cerca de 90% da produção mundial de café está concentrada meramente em 15 países, sendo que hoje o Brasil o maior produtor e exportador de café do mundo, com a produção de cerca de 45 milhões de sacas no ano de 2014 (CONAB, 2014). A área cultivada é aproximadamente de 2 milhões de hectares, porém com a grande quantidade de produtos que o país exporta, ele representa somente 6,9% das exportações do agronegócio (MAPA, 2014).

A produtividade vem crescendo com a adoção de tecnologias disponibilizadas por empresas de pesquisas. Esse aumento está relacionado com a introdução de novas cultivares, melhoramentos genéticos e disponibilização de informações relativas as cultivares, por exemplo: a indicação de cultivares adaptadas para determinada região do país.

## 2.2 Medidas Repetidas

Segundo Diggle (1988), a expressão medidas repetidas é definida com a observação de uma ou mais variáveis respostas na mesma unidade experimental (ou indivíduo) durante repetidas vezes. A natureza da resposta pode ser discreta ou contínua, multidimensional ou unidimensional.

De forma simples, a análise de medidas repetidas no tempo apresenta a possibilidade de investigar variações no comportamento de uma mesma unidade experimental (indivíduo) ou variações entre unidades ao longo do tempo. Destaca-se ainda a capacidade de obter um perfil médio para cada tratamento e um perfil individual de respostas para cada unidade.

A utilização de dados com medidas repetidas no tempo apresentam como vantagens o aumento na precisão da estimação da tendência da variável resposta ao longo do tempo (KUEHL, 2001).

Estudos de dados com medidas repetidas no tempo têm diversas áreas de aplicação como, por exemplo: na agronomia em análise de experimentos de plantas perenes, em estudos sobre eficiência do uso remédios e na biologia para indicar interações de uma doença com sistema imunológico.

Ao utilizar o estudo longitudinal conseguimos extrair informações das mudan-

ças ocorridas ao longo de um determinado tempo, porém ele é mais complexo que o estudo de dados transversais. Para essa análise devemos considerar os efeitos do tempo sobre as variáveis respostas e a possível correlação entre as respostas.

Existem diferentes técnicas para analisar os dados de um experimento com medidas repetidas, como a análise em parcelas subdivididas no tempo, que têm maiores restrições para a matriz de covariância e utilização de modelos mistos, que levam em conta a estrutura da matriz de covariâncias que melhor explicam os dados.

O esquema de parcelas subdivididas no tempo, em uma análise de medidas repetidas no tempo, por exemplo, deve considerar as parcelas como tratamentos e sub-parcelas como o tempo, porém os níveis do tempo não podem ser aleatorizados para seus intervalos. Assim, a análise de variância pode perder sua validade com a falta de aleatorização, pois os indivíduos podem apresentar uma matriz de covariância diferente daquela exigida em de uma análise usual, ou seja, variâncias homogêneas (XAVIER, 2000).

O uso da análise em parcelas subdivididas no tempo, na maioria das vezes não é a melhor opção para realizar a análise de dados com medidas repetidas. Esse modelo pressupõe que o erro da parcela, que engloba o fator de tratamentos e o erro da subparcela, onde são postos os tempos e a interação tempos x tratamento, tenham distribuição normal, sejam independentes, identicamente distribuídos. O erro da parcela é denominado como erro entre indivíduos, e o erro da subparcela como erro intra-indivíduos.

Segundo a autora Xavier (2000), uma condição suficiente para que seja válido o uso de parcelas subdivididas com medidas repetidas no tempo, é que a matriz de covariância  $\Sigma$  apresente a forma denominada de simetria composta, ou seja, sempre que for possível expressar  $\Sigma$  como:

$$\Sigma = \begin{bmatrix} (\sigma^2 + \sigma_1^2) & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & (\sigma^2 + \sigma_1^2) & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & (\sigma^2 + \sigma_1^2) & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & (\sigma^2 + \sigma_1^2) \end{bmatrix}$$

onde,

$\sigma^2$ : é a variância da subparcela (intra-indivíduos);

$\sigma_1^2$ : é a variância da parcela (entre indivíduos).

Essa simetria sugere que a variável aleatória esteja igualmente correlacionada e tenha a mesma variância, considerando medidas em diferentes instantes.

Caso a matriz de covariância não apresente a forma simétrica composta, o teste F será não exato e ocorrerá uma inflação sobre o efeito da subparcela. Pinheiro (1994) relatou que nesses casos, uma boa ferramenta para analisar seria o uso de modelos lineares mistos.

## 2.3 Modelos Lineares Mistos

Nas últimas décadas, devido à possibilidade de modelar a correlação intra-indivíduos os modelos lineares mistos tem sido um tema que desperta bastante interesse nos trabalhos acadêmicos. Considerando um experimento com medidas de um indivíduo repetidas no tempo, não podemos considerar que elas não são correlacionadas, sendo assim uma forma de modelar essa dependência é utilizar modelos lineares mistos (PINHEIRO, 1994).

Em uma análise com medidas repetidas no tempo, o modelo univariado em parcelas subdivididas no tempo necessita certa suposição na estrutura da matriz de covariância e caso não seja cumprida leva a erros nos testes. A alternativa é a análise com modelos mistos, pois eles permitem a utilização de várias estruturas de covariância no processo de modelagem, ou seja, escolher a que se adeque melhor aos dados.

Os modelos lineares de efeitos mistos incorporam a análise de curvas de crescimento e consideram a estrutura da matriz de covariâncias  $\Sigma$  que melhor explica as variações das observações, incluindo a vantagem de ajustar modelos com menor número de parâmetros (LITTELL et al. ,1998).

O modelo considera os efeitos fixos e aleatórios. A sua forma matricial é expressa como:

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i, \quad i = 1, \dots, N \quad (2.1)$$

onde:

N: é o número de unidades experimentais (indivíduos);

$Y_i$ : é um vetor  $(n_i \times 1)$  de medidas do i-ésimo indivíduo;



$n_i$ : é o número de observações no  $i$ -ésimo indivíduo;

$X_i$ : é uma matriz ( $n_i \times p$ ) que especifica os efeitos fixos;

$\beta$ : é um vetor ( $p \times 1$ ) parâmetros dos efeitos fixos;

$Z_i$ : é uma matriz ( $n_i \times q$ ) que especifica os efeitos aleatórios;

$b_i$ : é um vetor ( $q \times 1$ ) dos efeitos aleatórios;

$\epsilon_i$ : é um vetor ( $n_i \times 1$ ) de erros aleatórios.

No modelo é assumido que  $b_i$  e  $\epsilon_i$  são independentes, que  $b_i \sim N(0, D)$  e que  $\epsilon_i \sim N(0, R_i)$ . Com matrizes de covariâncias  $R_i(n_i \times n_i)$  e  $D(q \times q)$  representadas por

$$V \begin{bmatrix} b \\ \epsilon \end{bmatrix} = \begin{bmatrix} D & 0 \\ 0 & R \end{bmatrix}$$

A definição da matriz  $X$  é equivalente à utilizada na análise de regressão, ou seja, as colunas podem indicar os fatores que definem a estrutura dos efeitos fixos.

## 2.4 Estruturas das Matrizes de covariância

A utilização de modelos lineares mistos permite a seleção da melhor estrutura da matriz de covariância, ou seja, a que represente melhor a variabilidade dos dados. A escolha da estrutura de covariância influencia as estimativas e os erros padrões de efeitos fixos dessa forma pode-se obter diagnósticos e inferências equivocadas.

Algumas estruturas de matriz de covariâncias bastante utilizadas são:

1. Componente de Variância(VC):

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Tem variâncias iguais e observações independentes.

## 2. Simetria Composta (CS):

$$\Sigma = \begin{bmatrix} (\sigma^2 + \sigma_1^2) & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & (\sigma^2 + \sigma_1^2) & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & (\sigma^2 + \sigma_1^2) & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & (\sigma^2 + \sigma_1^2) \end{bmatrix}$$

Esta estrutura com variâncias iguais e covariâncias análogas entre medidas realizadas em diferentes instantes.

## 3. Sem estrutura (UN):

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

Esta estrutura permite que as variâncias e covariâncias sejam dessemelhantes. É útil para ser utilizada quando não existe o padrão de variâncias e covariâncias.

## 4. Auto-regressiva de 1ª ordem (AR(1)):

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

A correlação diminui à medida que aumenta o intervalo entre as medições, ou seja, a covariância também decresce. Em todos os instantes as variâncias são iguais. O parâmetro auto-regressivo é  $\rho$ , e no caso de um processo estacionário assume  $|\rho| < 1$ .

## 5. Auto-regressiva de 1ª Ordem Heterogênea (ARH(1)):

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

É a generalizada da estrutura Auto-regressiva de 1ª ordem, temos variâncias e covariâncias diferentes, sendo que o  $\rho$  é o parâmetro auto-regressivo satisfazendo  $|\rho| < 1$ .

#### 6. Auto-regressiva de primeira ordem com média móvel (ARMA(1,1))

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$$

Dados de séries temporais com parâmetro auto-regressivo  $\rho$ , componete de médias móveis *gamma*, sendo  $\sigma^2$  a variância residual.

#### 7. Toeplitz heterogênea

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 \\ \sigma_1\sigma_2\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 \\ \sigma_1\sigma_3\rho_2 & \sigma_2\sigma_3\rho_1 & \sigma_3^2 & \sigma_3\sigma_4\rho_1 \\ \sigma_1\sigma_4\rho_3 & \sigma_2\sigma_4\rho_2 & \sigma_3\sigma_4\rho_1 & \sigma_4^2 \end{bmatrix}$$

São dados de séres temporais igualmente espaçados, com parâmetros de variância diferentes para cada elemento da diagonal, sendo os elementos fora da diagonal principal funções de variâncias e do k-ésimo parâmetro de autocorrelação ( $\rho_k < 1$ ).

#### 8. Huynh-Feldt (HF):

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \frac{(\sigma_1^2 + \sigma_2^2)}{2} - \lambda & \frac{(\sigma_1^2 + \sigma_3^2)}{2} - \lambda & \frac{(\sigma_1^2 + \sigma_4^2)}{2} - \lambda \\ \frac{(\sigma_2^2 + \sigma_1^2)}{2} - \lambda & \sigma_2^2 & \frac{(\sigma_2^2 + \sigma_3^2)}{2} - \lambda & \frac{(\sigma_2^2 + \sigma_4^2)}{2} - \lambda \\ \frac{(\sigma_3^2 + \sigma_1^2)}{2} - \lambda & \frac{(\sigma_3^2 + \sigma_2^2)}{2} - \lambda & \sigma_3^2 & \frac{(\sigma_3^2 + \sigma_4^2)}{2} - \lambda \\ \frac{(\sigma_4^2 + \sigma_1^2)}{2} - \lambda & \frac{(\sigma_4^2 + \sigma_2^2)}{2} - \lambda & \frac{(\sigma_4^2 + \sigma_3^2)}{2} - \lambda & \sigma_4^2 \end{bmatrix}$$

Caracterizada por ter variâncias dessemelhantes nas diferentes ocasiões. As covariâncias são calculadas como a média aritimética entre as variâncias e subtraindo  $\lambda$ . O valor de  $\lambda$  é definido como a diferença entre a média das variâncias e a média das covariâncias.

9. Toeplitz (TOEP):

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

Essa estrutura é utilizada em séries temporais. Possui variâncias iguais e co-variâncias diferentes para situações distantes.

10. Estrutura Ante-Dependência de 1ª ordem:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 \\ \sigma_1\sigma_2\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_2 & \sigma_2\sigma_4\rho_2\rho_3 \\ \sigma_1\sigma_3\rho_1\rho_2 & \sigma_2\sigma_3\rho_2 & \sigma_3^2 & \sigma_3\sigma_4\rho_3 \\ \sigma_1\sigma_4\rho_1\rho_2\rho_3 & \sigma_2\sigma_4\rho_2\rho_3 & \sigma_3\sigma_4\rho_3 & \sigma_4^2 \end{bmatrix}$$

Parâmetros de variâncias diferentes para cada elemento da diagonal, onde os elementos fora da diagonal principal funções de variâncias e do k-ésimo parâmetro de auto-correlação, satisfazendo,  $|\rho| < 1$ .

Os termos  $\sigma^2$ ,  $\sigma$  e  $\rho$  correspondem, respectivamente, à variância, ao desvio-padrão e à correlação.

## 2.5 Estimação

Existem diferentes métodos para estimar os parâmetros  $\beta$  e  $b$  do modelo(2.1). Porém, há dois métodos bastante utilizados que são o Método de Máxima Verossimilhança (MV) e o de Máxima Verossimilhança Restrita (MVR).

Segundo o autor Searle et al. (1992), os estimadores de máxima verossimilhança apresentam características desejáveis como: suficiência, consistência e eficiência.

A sua forma clássica de estimação baseada na maximização da função verossimilhança marginal é dada por:

$$L_{MV}(\theta) = \prod_{i=1}^N (2\pi)^{\frac{n_i}{2}} |\Sigma_i(\alpha)|^{\frac{1}{2}} \exp \left( -\frac{1}{2} (y_i - X_i\beta)' \sum_i^{-1} (\alpha) (y_i - X_i\beta) \right)$$

onde,  $\alpha$  representa o vetor com todos os componentes de variância encontrado em  $\sum_i = Z_i D Z_i^T + R_i$ .

Os dois métodos são baseados no princípio da máxima verossimilhança. No entanto, podemos citar algumas diferenças como:

- os estimadores MVR não são tão sensíveis a *outliers* nos dados quando comparados a MV;
- os estimadores MVR não levam em consideração os graus de liberdade dos efeitos fixos do modelo;
- como as informações dos efeitos fixos ( $\beta$ ) não estão envolvidas no método MVR, os estimadores resultantes são invariantes ao valor de  $\beta$ .

O estimador de máxima verossimilhança (MV) de  $\beta$ , condicionada a  $\alpha$  é dado por:

$$\hat{\beta}(\alpha) = \left( \sum_{i=1}^N X_i' \sum_i^{-1} X_i \right)^{-1} \sum_{i=1}^N X_i' \sum_i^{-1} y_i$$

O método de máxima verossimilhança restrita (MVR) é uma derivação do método de máxima verossimilhança (MV).

## 2.6 Seleção do modelo

Diversos fatores podem influenciar a escolha de um modelo. Dessa forma, na seleção devemos levar em conta a parcimoniosidade, isto é, um modelo que envolva o mínimo de parâmetros possíveis a serem estimados e o ajuste do modelo escolhido.

A seleção dos modelos deve ser realizada com o auxílio de algum critério que justifique qual é o “melhor”. Por exemplo, um modelo que envolve uma quantidade excessiva de parâmetros poderá obter como resultado uma estimação ineficiente.

Os principais critérios de seleção de modelos são o critério de AKAIKE (Akaike's Information Criterion - AIC) e o bayesiano de Schwarz (Bayesian Information Criterion - BIC), ambos são baseados no valor da verossimilhança do modelo, dependem do número de parâmetros e do número de observações.

Os autores Emiliano (2009), Burnham (2004) apresentam os fundamentos e

aplicações para o melhor entendimento dos critérios de informação de Akaike e Bayesiano na seleção de modelos.

Os dois critérios são frequentemente utilizados para comparar modelos em diversas áreas, não existindo uma quantidade máxima de modelos a serem comparados. Ambos estão fundamentados na verossilhança, porém impõem diferentes penalizações. De acordo com esses critérios, o modelo que exibir o menor valor de AIC e BIC será o mais indicado.

### 2.6.1 Critério de informação de Akaike

O critério de informação Akaike - AIC foi elaborado por Hirotugu Akaike em 1971 como o nome de “um critério de informação” e proposto, em Akaike(1974). O critério é definido como:

$$AIC = -2\log L(\hat{\theta}) + 2(p),$$

onde  $L(\hat{\theta})$  é o estimador de máxima verossimilhança e  $p$  é o número de parâmetros no modelo.

O AIC tem a capacidade de avaliar a qualidade do modelo paramétrico, estimado pelo método da máxima verossimilhança. No caso de uma seleção entre vários modelos concorrentes, o valor do AIC obtido para cada modelo poderá ser utilizado em uma classificação, isto é, aqueles que apresentarem o menores valores do AIC são considerados os melhores. O critério não é um teste de hipóteses, não há significância e nem valor-p, porém com os valores do AIC pode-se concluir quais são os modelos mais adequados.

### 2.6.2 Critério de informação bayesiano

O critério de informação Bayesiano - BIC foi proposto por Schwarz(1978). O BIC é um critério que avalia modelos definidos em termos da probabilidade a posteriori, recebeu esse nome pois Schwarz deu um argumento Bayesiano para prová-lo. O BIC é calculado por:

$$BIC = -2 \log L(\hat{\theta}) + p \log N,$$

onde  $L\hat{\theta}$  é o estimador de máxima verossimilhança,  $p$  o número de parâmetros no modelo e  $N$  é o número de observações da amostra. Este critério penaliza modelos com muitos parâmetros. Na seleção do modelo mais adequado o escolhido será aquele que obtiver o menor valor do BIC.

No caso de modelos ajustados por máxima verossimilhança restrita (MVR) os valores de AIC, BIC e do teste de razão verossimilhança apenas podem ser comparados se os modelos possuírem a mesma estrutura de efeitos fixos. Porém, se os modelos comparados forem ajustados por máxima verossimilhança (MV), poderão ser comparados mesmo com estruturas de efeitos fixos diferentes.

Devemos selecionar o melhor modelo segundo os critérios AIC e BIC, porém estes critérios podem indicar diferentes modelos e então sugerir inferências parcialmente conflitantes. Caso exista essa incerteza na seleção, deve-se dar uma maior atenção para encontrar um modelo satisfatório para os dados.

### 2.6.3 Teste de Wald

Utiliza-se o teste de Wald para verificar a significância dos efeitos fixos de um modelo. As hipóteses em teste são:

$$H_0 : \beta_j = 0 \text{ versus } H_a : \beta_j \neq 0.$$

onde  $j$  representa o elemento do vetor de efeitos  $\beta$  em teste.

Para o testar as hipóteses, utiliza-se a matriz de covariância estimada dos efeitos estimados. O  $j$ -ésimo elemento da matriz é a variância estimada de  $\hat{\beta}_j$ ,  $\text{var}(\hat{\beta}_j)$ . A estatística do teste é dada por:

$$F_{1,gl} = \frac{\hat{\beta}_j^2}{\text{var}(\hat{\beta}_j)},$$

onde  $gl$  é o número de graus de liberdade.

A estatística de teste tem distribuição exata **F** com 1 grau de liberdade no numerador quando testa-se apenas um coeficiente. Outro teste para os parâmetros de covariâncias é o teste da razão de verossimilhança (TRV).

### 2.6.4 Teste da Razão de verossimilhança

O teste da razão de verossimilhança (TRV) é apropriado para comparar dois modelos hierárquicos ajustados por máxima verossimilhança (MV). O teste também pode ser aplicado em modelos ajustados por máxima verossimilhança restrita (MVR), porém ambos devem ser ajustados por MVR e ambos devem ter a mesma especificação dos efeitos fixos. Dessa forma, o TRV pode ser utilizado para testar hipóteses sobre parâmetros de efeito fixo ou parâmetros de covariância. A estatística  $2 \log(L_2/L_1)$  é baseado no logaritmo da razão entre as duas verossimilhanças dos modelos comparados (PINHEIRO & BATES, 2000). A estatística é calculada por:

$$2 \log(L_2/L_1) = 2[\log l(L_2) - \log l(L_1)] \sim \chi_g^2$$

onde,

$L_1$  : função de verossimilhança do modelo restrito;

$L_2$  : função de verossimilhança do modelo de referência.

Sob a hipótese  $H_0$  : o modelo restrito é adequado, o teste é assintótico e tem distribuição  $\chi^2$ , em que o número de graus de liberdade,  $g$ , é obtido subtraindo-se o número de parâmetros do modelo restrito do de parâmetros do modelo de referência. Caso a hipótese nula seja rejeitada, conclui-se que o modelo de referência é adequado. Em outras palavras, a rejeição de  $H_0$  indica que o modelo de referência com mais parâmetros apresenta um melhor ajuste.

Segundo o autor Matsushita (1994), uma desvantagem de utilizar o teste da razão de verossimilhança, é que somente 2 modelos podem ser comparados por vez, além do que um dos modelos é um caso especial do outro.



## 3 Material e Métodos

### 3.1 Material

O experimento foi instalado na região de Cerrado do Planalto Central em 2008, na área experimental da Embrapa Hortaliças, localizada na rodovia DF/158, Gama/DF. A área é caracterizada pelas seguintes coordenadas cartesianas e condições edafoclimáticas: latitude Sul: 15°56'00", longitude Oeste: 48°08'00", altitude: 997,62 m; área plana de Latossolo Vermelho Escuro com textura argilosa e apresentando média anual de 1600 mm de chuva, com duas estações típicas de períodos chuvosos e de seca, temperatura média anual de 22°C.

Os dados utilizados para a análise foram provenientes de um experimento de café realizado no período compreendido entre os anos 2010 a 2013. Um dos objetivos do experimento é verificar a resistência a ferrugem (praga) e adaptabilidade dos diversos cultivares de café. Durante os quatro anos foram avaliadas características das plantas como: altura, diâmetro do caule, projeção da copa, número de pares de ramos plagiotrópicos, produtividade, porém em nossa análise vamos considerar somente os dados de produtividade medida em sacas de 60 kg de café beneficiado por hectare (sc/ha).

O delineamento experimental utilizado foi de blocos ao acaso, com 4 blocos, contendo 10 plantas por parcela. O material experimental consistiu de 30 cultivares de café (tratamentos), totalizando a quantidade de 1200 plantas no experimento. As cultivares e progênies de café foram desenvolvidas por instituições, sendo que 23 pertencem ao grupo de cultivares resistentes à ferrugem, 4 progênies experimentais e 3 suscetíveis à ferrugem. As médias e desvios padrões da produtividade do café por cultivar, conforme a Tabela 3.1 a seguir.

Tabela 3.1: Relação das cultivares de café com suas respectivas médias e desvios padrões (D.P) da produtividade .

Tratamento	Cultivar	Instituição	Prod. M (sc/ha)	D.P
1	Catuaí Amarelo 2SL	PROCAFÉ	62,20	29,44
2	Catuaí Amarelo 24/137	PROCAFÉ	50,94	24,75
3	Catuaí Amarelo 10/15 cv479	PROCAFÉ	56,18	38,82
4	Catuaí Vermelho 785/15	PROCAFÉ	42,33	21,43
5	Catuaí Vermelho 20/15 cv476	PROCAFÉ	54,13	28,85
6	Sabiá 398	PROCAFÉ	61,02	48,51
7	Palma II	PROCAFÉ	68,82	48,58
8	Acauã	PROCAFÉ	70,86	29,47
9	Oeiras MG 6851	EPAMIG	59,65	26,57
10	Catiguá MG 1	EPAMIG	42,82	13,83
11	Sacramento MG 1	EPAMIG	48,25	16,60
12	Catiguá MG 2	EPAMIG	51,61	18,58
13	Araponga MG 1	EPAMIG	67,45	40,98
14	Paraíso MG 419-1	EPAMIG	47,26	17,79
15	Pau Brasil MG 1	EPAMIG	49,60	24,30
16	Tupi IAC 1669-33	IAC	65,01	19,90
17	Obatã Vermelho IAC 1669-20	IAC	68,78	38,60
18	IAPAR 59	IAPAR	56,49	27,83
19	IPR 98	IAPAR	62,13	25,67
20	IPR 99	IAPAR	55,81	33,72
21	IPR 103	IAPAR	70,75	35,84
22	IPR 104	IAPAR	55,20	26,62
23	Catiguá MG 3	EPAMIG	47,65	20,07
24	Topázio MG 1190	EPAMIG	51,08	34,52
25	Catuaí Vermelho IAC 144	IAC	55,42	28,03
26	H419-3-3-7-16-4-1	EPAMIG	59,99	31,31
27	H419-10-6-2-5-1	EPAMIG	56,04	24,60
28	H419-10-6-2-10-1	EPAMIG	45,35	21,30
29	H419-10-6-2-12-1	EPAMIG	51,02	28,64
30	Catuaí Amarelo IAC 62	IAC	49,65	24,63

Para o presente trabalho será utilizada a variável produtividade, para qual foram realizadas 4 medidas ao longo do tempo, onde: A0, A1, A2 e A3 foram as medidas realizadas nos anos de 2010, 2011, 2012 e 2013 respectivamente.

São 480 observações referentes a variável produtividade, dessa maneira serão apresentados apenas 15 observações na Tabela 3.2.

Tabela 3.2: Dados obtidos do experimento de café em função dos tratamentos e medidas realizadas ao longo do tempo para a variável produtividade.

Ano (A)	Tratamento (T)	Bloco (B)	Produtividade (P) (sc/ha)
2013	1	1	63,90
2013	1	2	46,00
2013	1	3	30,30
2013	1	4	36,20
2013	2	1	16,40
2013	2	2	40,10
2013	2	3	30,20
2013	2	4	24,80
2013	3	1	11,10
2013	3	2	8,50
2013	3	3	45,20
2013	3	4	24,80
2013	4	1	31,00
2013	4	2	22,40
2013	4	3	27,10

Os 30 perfis individuais de produtividade do café foram separados por cultivares estão representados na Figura 3.1.

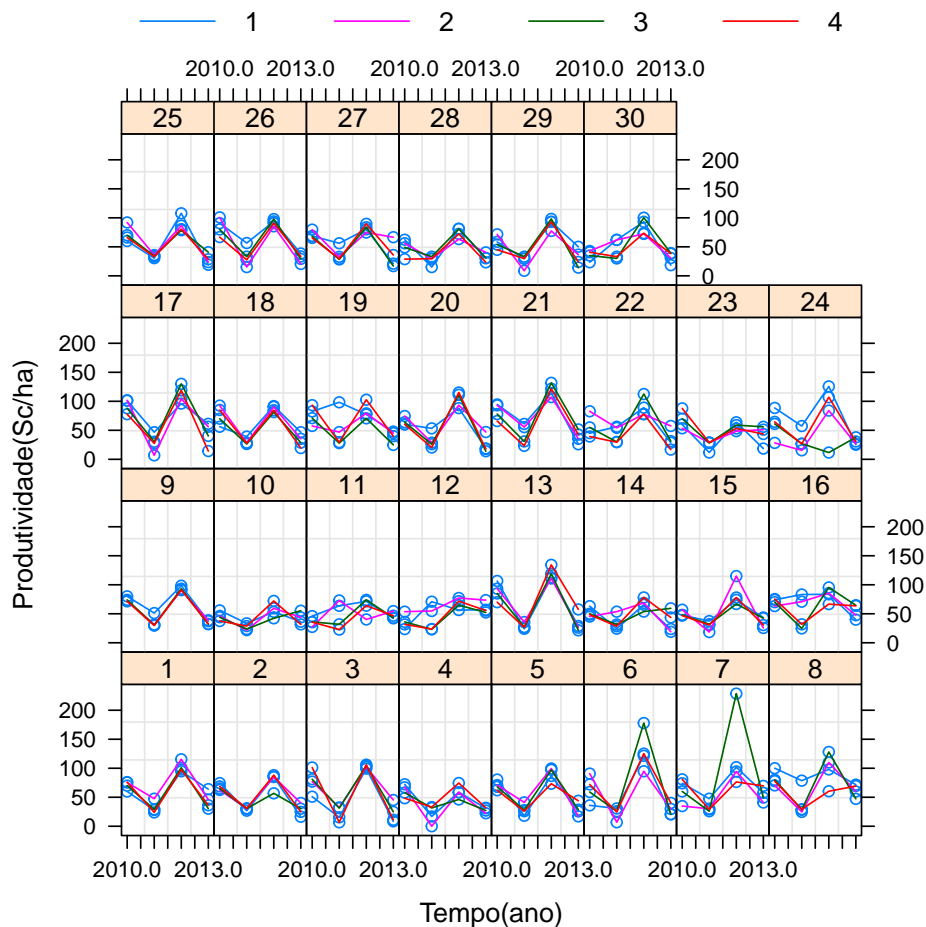


Figura 3.1: Perfis individuais da produtividade do café, por cultivar.

Pode-se observar na Figura 3.1 uma variabilidade na produtividade do café nos anos em cada indivíduo. Verifica-se ainda a presença de um ciclo bienal alternando entre alta e baixa produtividade. Em relação a algumas cultivares específicas nota-se uma menor variabilidade produtividade. Temos como exemplo as cultivares 10 e 11.

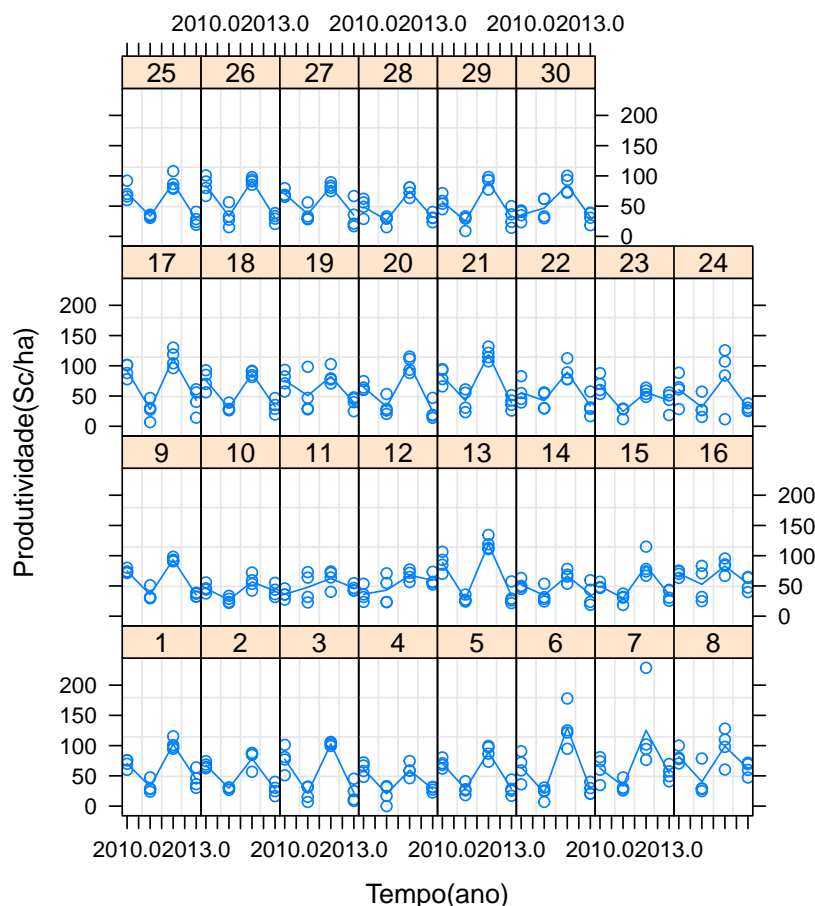


Figura 3.2: Perfis médios da produtividade do café, por cultivar.

## 3.2 Métodos

O processo para a construção de um modelo linear misto pode ser descrito em poucos passos:

- Analisar os gráficos de perfis individuais e médios, que poderá sugerir a introdução de um efeito aleatório em algum dos coeficientes.
- Escolher os efeitos fixos e aleatórios com o uso de critérios de informação para a seleção.
- Escolher as estruturas de covariâncias entre indivíduos  $\mathbf{D}$  e intra indivíduos  $\mathbf{R}_i$  que melhor expliquem as variações dos dados.
- Realizar o diagnóstico para verificar a qualidade do ajuste do modelo e a validade das suposições de normalidade dos resíduos e heterogeneidade de variâncias.

O procedimento para a escolha dos efeitos fixos e aleatórios que farão parte do modelo misto, inicia-se com a análise dos gráficos de perfis individuais das respostas para cada nível do fator longitudinal. Este gráfico poderá sugerir a inclusão de efeito aleatório em algum dos coeficientes. O modelo inicial teve um efeito aleatório associado ao intercepto e ao efeito de ano.

Após a análise dos gráficos de perfis individuais na Figura 3.1 e médios na Figura 3.2, presume-se que a produtividade do café do  $i$ -ésimo indivíduo medida no tempo  $a_{ij}$  pode ser explicada pelo modelo:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 cult_i + \beta_2 ano_j + \beta_3 bloco_j + \beta_4 cult_i ano_j}_{ef. fixos} + \underbrace{b_{0i} + b_{1i} ano_{ij} + \epsilon_{ij}}_{ef. aleatórios} \quad (3.1)$$

Onde,  $i = 1, \dots, 120$  e  $j = 1, \dots, 4$ .

O modelo inicial inclui os efeitos de cultivar, ano, bloco e interação entre cultivar e ano. Temos também os efeitos aleatórios relacionados a cada indivíduo sendo eles o intercepto e o efeito de ano.

Durante a seleção dos efeitos que comporão o modelo misto, será utilizado o padrão do *software* R, admitindo que a matriz de covariâncias entre indivíduos  $\mathbf{D}$  tem uma estrutura UN (não estruturada) e que a matriz de covariâncias intra-indivíduos  $\mathbf{R}_i$  tem uma estrutura VC (componente de variância).

Depois de selecionada a estrutura aleatória do modelo misto, iniciam-se as comparações com outros modelos encaixados utilizando os resultados dos Critérios de Informação Akaike (AIC), Critério de Informação Bayesiano (BIC) e o teste de razão verossimilhança para a escolha do melhor modelo.

A seleção dos efeitos fixos é finalizada no momento que um modelo parcimonioso, que envolva o mínimo de parâmetros possíveis e que explique bem o comportamento do fenômeno em estudo for encontrado. Além dos critérios de informação, pode-se também verificar os efeitos fixos significativos, através do teste de Wald.

Após a escolha dos efeitos fixos e aleatórios que comporão o modelo deve-se especificar uma estrutura para a matriz de covariâncias entre indivíduos  $\mathbf{D}$ , admitindo que  $\mathbf{R}_i = \sigma^2 \mathbf{I}$  conhecida como Componente de Variância (VC).

Os parâmetros  $\beta_0, \beta_1, \beta_2, \beta_3$  e  $\beta_4$  representam respectivamente os efeitos fixos relativos ao intercepto, às variáveis indicadoras de cultivar, ano, bloco e a interação entre cultivar e ano. Da mesma maneira, os termos  $b_0$  e  $b_1$  constituem os efeitos aleatórios relativos ao intercepto e efeito de ano.

O vetor de efeitos aleatórios é dado por  $b_i = (b_{0,i}, b_{1,i})^T \sim N(0, \mathbf{D})$ , onde

$$\mathbf{D} = \begin{pmatrix} \sigma_{b_0}^2 & \sigma_{b_0 a} \\ \sigma_{b_0 a} & \sigma_a^2 \end{pmatrix}$$

Cada observação  $y_{ij}$  possui um erro  $e_{ij}$  associado, que tem distribuição normal  $e_i \sim N(0, \mathbf{R}_i)$ , onde

$$\mathbf{R}_i = \begin{pmatrix} Var(e_{i1}) & cov(e_{i1}, e_{i2}) & \dots & Cov(e_{i1}, e_{i4}) \\ cov(e_{i1}, e_{i2}) & Var(e_{i2}) & \dots & Cov(e_{i2}, e_{i4}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(e_{i1}, e_{i4}) & cov(e_{i2}, e_{i4}) & \dots & Var(e_{i4}) \end{pmatrix}$$

As estruturas das matrizes de covariância a serem testadas foram descritas na Seção 2.4. Porém foram utilizadas apenas as estruturas diagonal, não estruturada, identidade e simetria composta.

O diagnóstico do modelo ajustado é realizado principalmente através de métodos gráficos. Um bom ajuste é obtido quando os pontos do gráfico dos valores preditos versus valores observados estão próximos de uma reta com inclinação de  $45^\circ$  passando pela origem do gráfico.

A suposição de normalidade dos resíduos e heterogeneidade de variâncias pode ser verificada por meio do gráfico quantil-quantil da Normal e pelo histograma. Pode-se também verificar essa suposição realizando o teste Shapiro-Wilk e Kolmogorov-Smirnov.

## 4 Resultados e Discussão

Durante a avaliação do gráfico de boxplot da produtividade do café por cultivar apresentado na Figura 4.1, verifica-se a presença de cultivares com alta produtividade, porém com alta variabilidade. Também verifica-se a presença de candidatos a outlier nas cultivares 7 e 15. Caso exerçam influências desproporcionais tais observações poderão ser retiradas posteriormente.

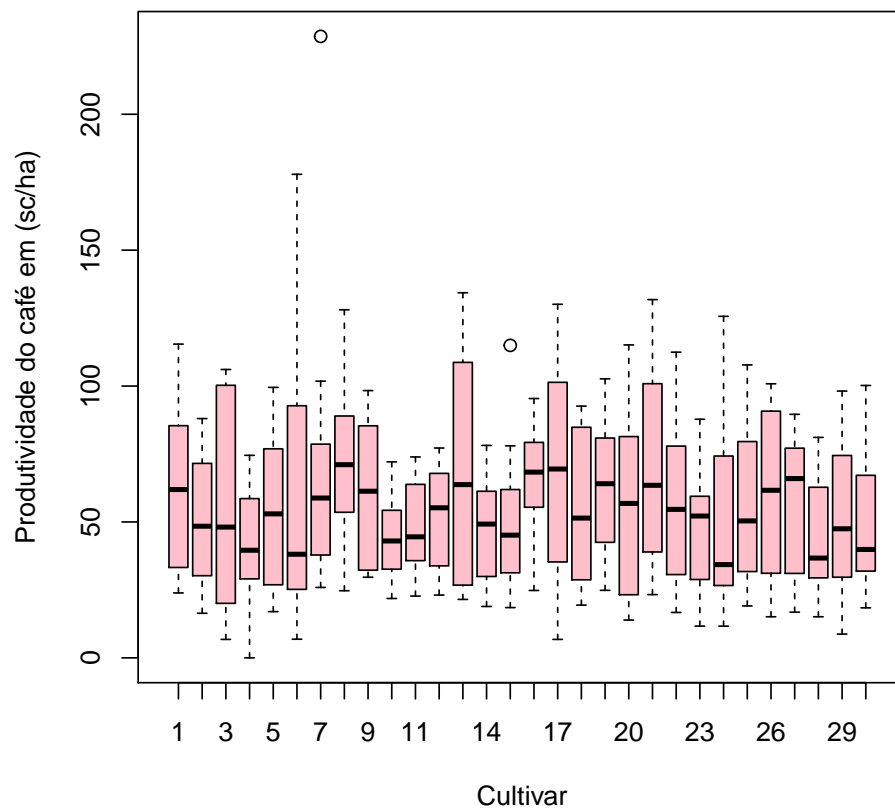


Figura 4.1: Boxplot da produtividade do café em sacas por hectare (sc/ha), por cultivar.

Na Figura 4.2, percebe-se que a produtividade do café apresenta um ciclo de alta produtividade seguido de uma baixa. O fenômeno de queda na produtividade foi explicado por Mendonça et al.(2011), devido ao desgaste causado nas plantas nos anos de alta produtividade. Verifica-se, ainda, que a variabilidade nos anos parece ser pequena,



porém há uma grande variação na produtividade entre os anos reforçando a existência do fenômeno de bienalidade.

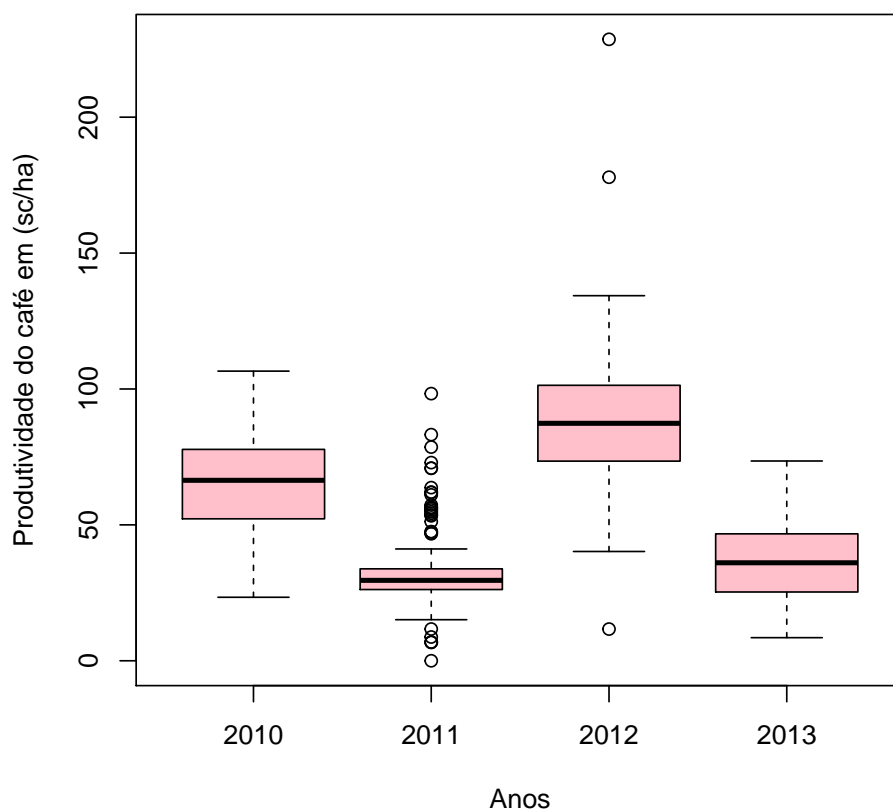


Figura 4.2: Boxplot da produtividade do café em sacas por hectare (sc/ha), por ano.

De acordo com os dados da Tabela 4.1, tem-se uma produtividade do café média global de 56,11 sc/ha, variando de 0,00 sc/ha a 228,65 sc/ha. Certifica-se que apesar da existência do ciclo produtivo de alta seguido de baixa, houve um aumento na produtividade média do segundo biênio.

Tabela 4.1: Valores médios, desvios padrão, mínimos, máximos de produtividade do café nos quatro anos.

	Anos				Geral
	2010	2011	2012	2013	
Média	65,15	33,64	88,69	36,97	56,11
D.P	18,98	16,08	26,50	14,91	
Mínimo	23,36	0,00	11,66	8,50	
Máximo	106,57	98,29	228,65	73,50	

Na Tabela abaixo, tem-se as estimativas das correlações, variâncias e covariâncias entre os anos 2010, 2011, 2012 e 2013. Essa tabela pode apontar a necessidade do uso de modelos lineares mistos.

Tabela 4.2: Correlações (abaixo da diagonal), variâncias (diagonal) e covariâncias (acima da diagonal) amostrais, da produtividade do café, por ano.

	2010	2011	2012	2013
2010	360,42	-12,47	152,71	7,27
2011	<b>-0,04</b>	258,85	-27,25	51,58
2012	<b>0,30</b>	<b>-0,06</b>	702,25	-58,14
2013	<b>0,02</b>	<b>0,21</b>	<b>-0,14</b>	222,52

Encontram-se evidências para utilizar modelos mistos, dado que as variâncias aparentemente são desiguais, pois são maiores em anos com alta produtividade. Da mesma forma, verifica-se que existe correlação não nula entre os anos 2010-2012 (anos com alta produtividade) e 2011-2013 (anos com baixa produtividade). Esses são aspectos que indicam a necessidade do uso de diferentes estruturas através de modelos lineares mistos.

Nos gráficos de perfis individuais de produtividade do café exibidos na Figura 3.1, também pode-se observar alguns desses aspectos, que sugerem a inclusão de efeito aleatório.

Na escolha da estrutura dos efeitos aleatórios, pode-se identificar fontes de variação nos valores medidos, sendo elas devido ao bloco, cultivar, ano, à variabilidade

entre indivíduos e devida à variabilidade das medidas realizadas nos indivíduos. O efeito aleatório verificado no gráfico de perfis individuais representa a variabilidade de medidas fornecidas por diferentes indivíduos (entre indivíduos) do mesmo cultivar. Para essa análise, será considerado os efeitos fixos de bloco, ano e cultivar.

O processo de escolha da estrutura fixa iniciou-se com uma estrutura maximal. A comparação entre os modelos foi realizada com o auxílio dos critérios AIC, BIC e do teste de razão verossimilhança (TRV).

Na Tabela 4.3 estão as estruturas de efeito fixos que foram comparadas. Os resultados dessas comparações se encontram na Tabela 4.4.

Tabela 4.3: Estruturas dos efeitos fixos, aleatórios e seu método de estimação, para dados de produtividade do café.

Modelo	Método de Estimação	Efeito Fixo	Efeito Aleatório
M1	MVR	cult., ano, bloco e cult. vs. ano	intercepto e ano
M1.1	MV	cult., ano, bloco e cult. vs. ano	intercepto e ano
M2	MVR	cult., ano, bloco	intercepto e ano
M2.1	MV	cult., ano, bloco	intercepto e ano
M3	MRV	cult., ano	intercepto e ano
M3.1	MV	cult., ano	intercepto e ano

Nota: MVR- Máxima Verossimilhança Restrita e MV- Máxima Verossimilhança

Tabela 4.4: Comparação dos modelos M1.1, M2.1 e M3.1 pelos critérios de informação AIC e BIC e pelo teste da razão verossimilhança (TRV)

Modelo	gl	AIC	BIC	-2log Veross	Comparação	$\chi^2_{calc}$	p-valor
M1.1	134	4091.47	4650.76	-1911.73			
M2.1	47	4134.55	4330.71	-2020.27	M1.1 vs M2.1	217.07	<.0001
M3.1	44	4138.75	4322.39	-2025.37	M2.1 vs M3.1	10.20	0.0169

Os resultados das comparações apresentadas na Tabela 4.4 demonstram que o modelo M1.1 pelo critério AIC apresentou o melhor ajuste. Porém pelo critério BIC que

penaliza o modelo com maior número de parâmetros, o modelo M3.1 se ajustou melhor. Considerou-se o modelo M1.1, dado que efeito da interação cultivar vs. ano é significativo.

Para verificar a significância dos coeficientes do modelo M1.1 utilizou-se o teste de Wald, com o modelo ajustado pelo método de máxima verossimilhança (MV). Com resultados da Tabela 4.5 pode-se concluir que os todos os coeficientes do modelo M1.1 são significativamente ( $p < 0,05$ ) não nulos.

Tabela 4.5: Teste de Wald para os parâmetros de efeito fixo do modelo M1.1

Coeficiente	gl. numerador	gl. denominador	<b>F</b>	p-valor
Intercepto	1	270	6528.66	<.0001
Ano	3	270	257.698	<.0001
Cultivar	29	87	4.445	<.0001
Bloco	3	87	3.306	0.0239
Cult:Ano	87	270	2.938	<.0001

Após definir os efeitos aleatórios do modelo, foi selecionada a estrutura de covariâncias para os efeitos aleatórios **D**. Então, realizou-se a comparação entre a estrutura diagonal (Diag) mais simples com a não estruturada (UN), respectivamente os modelos M1.3 e M1.2. Para esta fase utilizou-se a estrutura de covariância intra-indivíduo  $\mathbf{R} = \mathbf{I}\sigma^2$ .

Tabela 4.6: Comparação dos modelos M1.2, M1.3 pelo teste da razão verossimilhança (TRV)

Modelo	Estrutura ( <b>D</b> )	gl	AIC	BIC	-2log	Comparação	$\chi^2_{calc}$	p-valor
M1.2	UN	134	4091.47	4650.76	-1911.73			
M1.3	Diag	128	4087.93	4622.18	-1915.97	M1.2 vs M1.3	8.46	0.206

Ao avaliar os resultados da Tabela 4.6, verifica-se que o modelo M1.3 com menos parâmetros não é rejeitado pelo teste TRV e apresenta o menor valor de AIC. Logo, a estrutura diagonal representa bem os dados, pois assumindo essa estrutura para a matriz de covariâncias **D** admite-se independência entre os interceptos e efeito aleatório de ano. Nessa etapa escolhe-se o modelo M1.3 (diagonal).

Para auxiliar a escolha de uma estrutura de covariâncias intra-indivíduos diferente do padrão sugerido pelo *software*,  $\mathbf{R} = \mathbf{I}\sigma$ , utiliza-se o gráfico de perfis individuais como referência.

Tabela 4.7: Comparação dos modelos M1.4, M1.5 e M1.6 pelos critérios AIC, BIC e teste da razão verossimilhança (TRV)

Modelo	Estrutura ( $\mathbf{R}_i$ )	gl	AIC	BIC	-2log	Comparação	$\chi^2_{calc}$	p-valor
M1.4	UN	134	3397.02	3916.64	-1564.51			
M1.5	CS	129	3393.22	3893.45	-1567.61	M1.4 vs M1.5	6.20	0.287
M1.6	AR1	129	3390.90	3891.13	-1566.45	M1.4 vs M1.6	3.87	0.567

Dentre as estruturas para matriz de covariâncias ( $\mathbf{R}_i$ ) testadas, pelo teste da razão de verossimilhança as comparações não foram favoráveis a escolha do modelo M1.4 sem estrutura (UN).

Apesar do modelo M1.4 não ser o indicado pelos critério de seleção, entre os testados é o único que apresenta um número satisfatório de parâmetros para explicar as correlações existentes entre anos com alta e baixa produtividade. Os modelos M1.5 e M1.6 estimam apenas 1 parâmetro de correlação, o que não condiz com estrutura das correlações existente.

Ao final dessa etapa de escolha da estrutura de covariâncias intra-indivíduo ( $\mathbf{R}_i$ ) que representa bem a variabilidade das medidas, escolheu-se o modelo M1.4.

Com as estimativas dos parâmetros do modelo M1.4, pode-se obter a equação que descreve a produtividade. Como exemplo temos o indivíduo 7 no ano de 2011 com valor observado de  $y_{7,2} = 30,43$ . Seu valor estimado pelo modelo M1.4 seria  $\hat{y}_{7,2} = 28,40$  pela equação:

$$\hat{y}_{7,2} = 73.55 - 2.71 - 38.26 - 4.33 + 0.15$$

Como um dos objetivos do experimento é encontrar a cultivar que melhor se adapta as condições climáticas da região, ou seja, com maior produtividade. Existe a necessidade de verificar se existe uma diferença significativa entre as 4 cultivares com

maior produtividade média de sacas de café beneficiado por hectare (sc/ha). Realizando o teste de comparação de médias Tukey obtemos os resultados da tabela 4.8 a seguir:

Tabela 4.8: Resultados do teste de Tukey para comparação de médias das cultivares com maior produtividade média

Cultivar	Produtividade Média (sc/ha)
Acauã (8)	70.86 a
IPR 103 (21)	70.75 ab
Palma II (7)	68.82
Obatã vermelho IAC (17)	68.78 b

Temos que, médias de produtividade seguidas de letras iguais não diferem significativamente, pelo teste de Tukey ao nível de probabilidade de 5%.

Dessa maneira, pode-se concluir que os cultivares acauã (8) e IPR(21) não apresentam diferenças significativas com relação a produtividade média. Sendo que, estas cultivares foram as únicas a apresentar produtividades médias superiores a 70 (sc/ha).

A análise gráfica tem com objetivo de verificar a qualidade do ajuste do modelo M1.4 e a validade dos pressupostos de normalidade, homocedasticidade dos resíduos e aditividade.

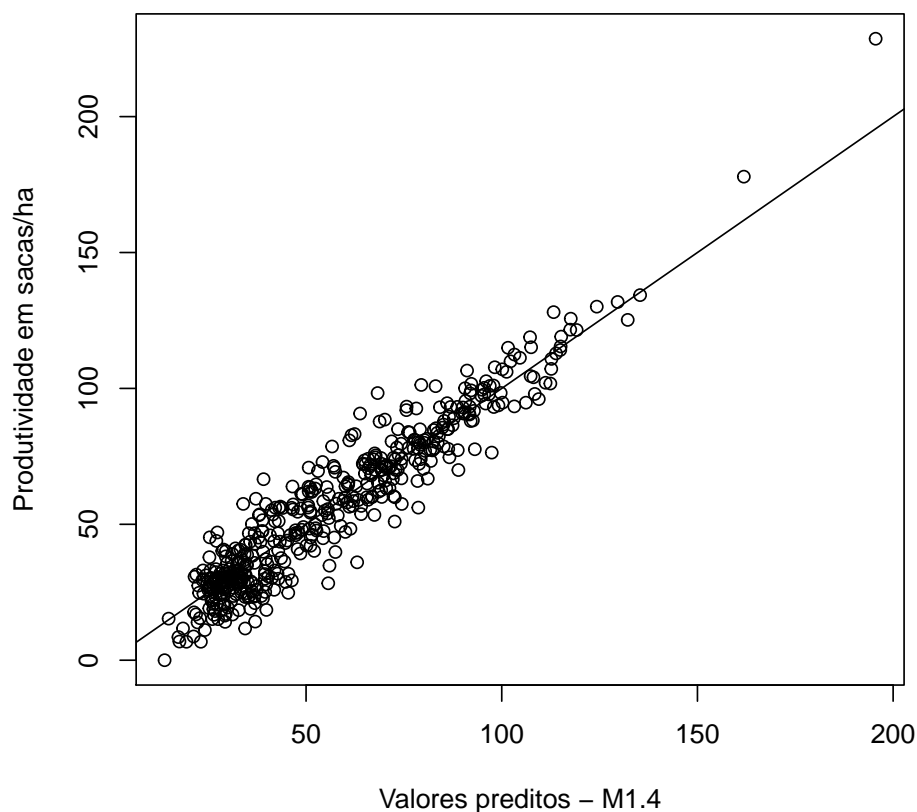


Figura 4.3: Ajuste do modelo M1.4 aos dados de produtividade do café

Na Figura 4.3, verifica-se que os valores preditos estão próximos a reta, o que indica que os valores ajustados estão próximos aos valores observados, confirmando o bom ajuste do modelo M1.4 aos dados.

Pode-se verificar nas figuras abaixo os gráficos de diagnóstico dos resíduos. Apesar da existência de pontos atípicos, na Figura 4.4 verifica-se que o modelo M1.4 foi capaz de se acomodar bem aos dados, visto que nenhum dos pontos ficaram fora da faixa de -3 a 3.

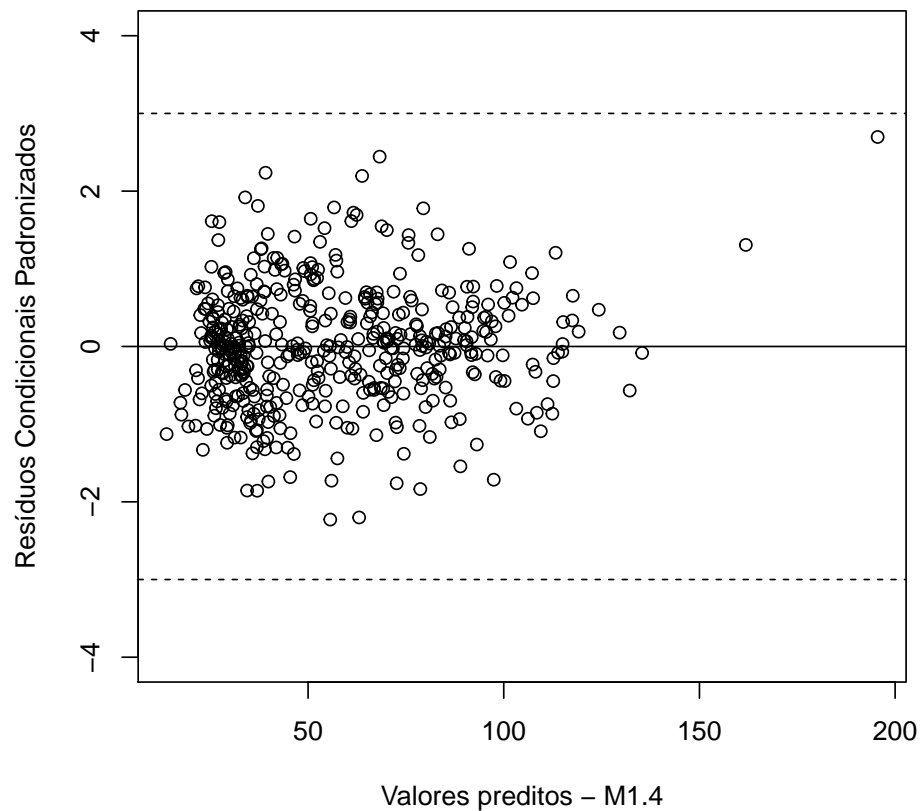
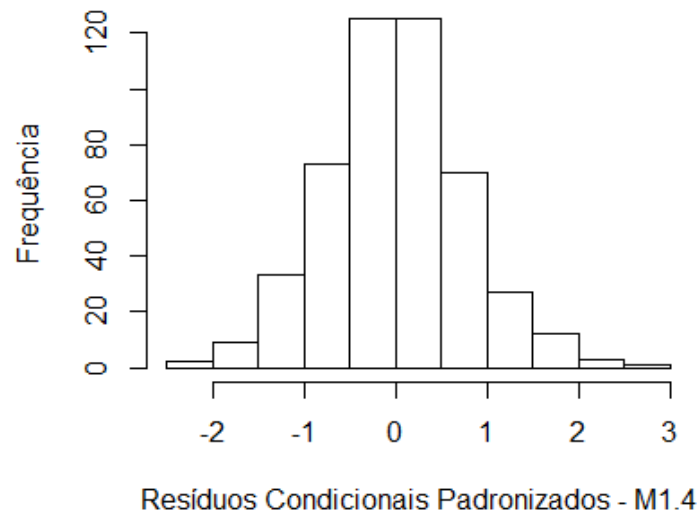


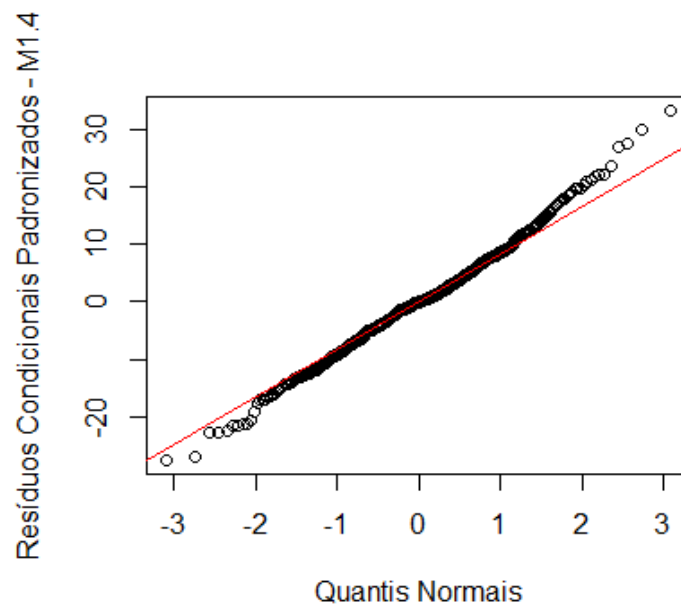
Figura 4.4: Gráfico de diagnóstico dos resíduos condicionais padronizados

Com relação a normalidade, pode-se observar na Figura 4.5 (a) que os resíduos do modelo M1.4 se aproximam de uma distribuição normal, pois se agrupam em torno da linha gerada no diagrama quantil-quantil. Realizando o teste de normalidade Shapiro-wilk encontramos um valor de  $p = 0.056$ , sob hipótese nula de que os dados seguem uma distribuição normal.





(a)



(b)

Figura 4.5: Gráficos de diagnóstico dos resíduos

O histograma na Figura 4.5(a) apresenta o formato característico da distribuição normal, uma vez que observa-se simetria ao redor do zero, com frequência mais alta no centro diminuindo conforme se aproxima das bordas. Apesar do histograma possuir um peso maior nas caudas, todas as evidências apontam para a normalidade dos resíduos.

## 5 Considerações finais

Com base nos resultados apresentados podemos realizar as seguintes considerações:

A utilização de modelos lineares mistos para a análise dos dados de produtividade do café de um experimento com medidas repetidas no tempo, foi bem sucedida, tendo em vista que o modelo apresentou um bom ajuste aos dados, representando satisfatoriamente a sua variabilidade.

As cultivares com as maiores produtividades de grãos, ou seja, mais adaptadas às condições ambientais foram: Acauã (8) e IPR 103 (21), com produtividades médias acima de 70 sacas por hectare de café beneficiado. Porém, ao realizar o teste de comparação de médias, concluímos que as cultivares Acauã (8) e IPR 103 (21), a um nível de significância de 5%, não diferem quanto a produtividade.

Com o passar do tempo com um maior número de observações geradas pelo experimento, o modelo tende a se tornar mais robusto e com menores níveis ruídos.

Com relação as estruturas de covariâncias intra-indivíduos mais estudos devem ser realizados com a adição de novas observações.

## Referências Bibliográficas

- [1] ABIC - ASSOCIAÇÃO BRASILEIRA DE INDÚSTRIA DE CAFÉ. A história do café. Disponível em: <http://www.abic.com.br/publicar/cgi/cgilua.exe/sys/start.htm?sid=55>, 2014. Acesso em 5 jul. 2015.
- [2] AKAIKE, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 19, 6 (1974), 716–723.
- [3] ALCARDE, R. *Modelos lineares mistos em dados longitudinais com o uso do pacote ASReml-R*. Tese (doutorado em estatística e experimentação agrônômica), Escola superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2012.
- [4] ARTERO, A. S. *Caracterização Citomolecular da Espécie Coffea Arábica L*. Dissertação (mestrado em agricultura tropical e subtropical), Instituto Agronômico, Campinas, São Paulo, 2006.
- [5] BARBOSA, M. *Uma abordagem para análise de dados com medidas repetidas utilizando modelos lineares mistos*. Dissertação (mestrado em estatística e experimentação agrônômica), Escola superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2009.
- [6] BOX, G. E., ET AL. Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *The annals of mathematical statistics* 25, 2 (1954), 290–302.
- [7] BURNHAM, K. P., AND ANDERSON, D. R. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research* 33, 2 (2004), 261–304.
- [8] CAMARINHA FILHO, J. A. *Modelos lineares mistos: estruturas de matrizes de variâncias e covariâncias e seleção de modelos*. Tese (doutorado em estatística

- e experimentação agrônômica), Escola superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2002.
- [9] CAMARINHA FILHO, J. A. *Nota Metodológica sobre Modelos Lineares Mistos*. Notas de aula, Universidade Federal do Paraná, Departamento de Estatística, 2002.
- [10] DE ABASTECIMENTO, C. C. N. Série histórica de Área em produção de café. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/cafetotalseriehist.xls>, 2014. Acesso em 19 jun. 2015.
- [11] DE FREITAS, E. G., ET AL. Modelo univariado aplicado a dados longitudinais de cana de açúcar. *Revista Brasileira de Biometria, São Paulo* 26, 2 (2008), 93–106.
- [12] EMILIANO, P. C. *Fundamentos e aplicações dos critérios de Informação: akaike e bayesiano*. Dissertação (mestrado em estatística e experimentação), Universidade Federal de Lavras, Minas Gerais, 2009.
- [13] FARAWAY, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2005.
- [14] FERREIRA, W. L. *Análise de dados com medidas repetidas em experimento com ingestão de café*. Dissertação (mestrado em estatística e experimentação agropecuária), Universidade Federal de Lavras, Minas Gerais, 2012.
- [15] GEISSER, S., GREENHOUSE, S. W., ET AL. An extension of box’s results on the use of the  $f$  distribution in multivariate analysis. *The Annals of Mathematical Statistics* 29, 3 (1958), 885–891.
- [16] KUEHL, R. O., AND OSUNA, M. G. *Diseño de experimentos: Principios estadísticos de diseño y análisis de investigación*, vol. 2. International Thomson México, 2001.
- [17] LAIRD, N. M., AND WARE, J. H. Random-effects models for longitudinal data. *Biometrics* 38, 4 (1982), 963–974.
- [18] MALHEIROS, E. B. Precisão da análise de experimentos com medidas repetidas no tempo usando procedimentos do sas. *Revista de matemática e estatística* 19 (2001), 253–272.

- [19] MATSUSHITA, R. Y. *Modelos Longitudinais Mistos com Correlação Serial nos Erros*. Dissertação (mestrado em estatística), Universidade Estadual de Campinas, São Paulo, 1994.
- [20] MENDONÇA, R., RODRIGUES, W., MARTINS, L., AND TOMAZ, M. Abordagem sobre a bienalidade de produção em plantas de café. *Enciclopédia Biosfera* 7, 13 (2011), 1–9.
- [21] MONTGOMERY, D. C. *Design and analysis of experiments*, 8 ed., vol. 7. Wiley New York, 1984.
- [22] PINHEIRO, J., AND BATES, D. *Mixed-effects models in S and S-PLUS*, 3 ed. Springer Science & Business Media, 2006.
- [23] PINHEIRO, J. C. *Topics in mixed effects models*. Tese (doutorado em estatística), University of Wisconsin, Madison, 1994.
- [24] SEARLE, S. R., CASELLA, G., AND MCCULLOCH, C. E. *Variance components*, vol. 391. John Wiley & Sons, 2009.
- [25] SOUZA, T. G. *Tópicos em análise de experimentos longitudinais para aplicações em estudos de sinais biopotenciais*. Dissertação (mestrado em estatística), Universidade de Brasília, 2013.
- [26] VEIGA, A. D., GUERRA, A. F., ROCHA, O. C., BARTHOLO, G. F., RODRIGUES, G. C., AMABILE, R. F., AND SILVA, T. P. D. *Avaliação de genótipos de café arábica resistentes à ferrugem no Cerrado do Planalto Central*. IX simpósio de pesquisa dos cafés do brasil, Curitiba-PR, 2015.
- [27] VIVALDI, L. J. *Análise de experimentos com dados repetidos ao longo do tempo ou espaço*. Embrapa Cerrados, Planaltina, Distrito Federal, 1999.
- [28] WOLFINGER, R., AND CHANG, M. *Comparing the SAS GLM and MIXED procedures for repeated measures*. Citeseer, 1998.
- [29] XAVIER, L. H. *Modelos univariados e multivariados para análise de medidas repetidas e verificação da acurácia do modelo univariado por meio de simulação*. Dissertação (mestrado em estatística e experimentação agrônômica), Escola superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2000.

# Apêndices

# A Programa no Software R para o ajuste do modelo linear misto

```
#Pacotes:
require(nlme)
require(lattice)
require(car)
require(multcomp)

# Renomeando as variáveis:
dados<-dados.anos.x.prod2 <- read.table("C:/Users/Sony/Dropbox/Pen Drive
/TCC1/Dados/dados anos x prod2.txt", header=TRUE)
ano<-as.factor(dados$Ano)
ano.n<-dados$Ano
cult<-as.factor(dados$Cultivar)
bloco<-as.factor(dados$Bloco)
levels(bloco)<-c("A","B","C","D")
prod<-dados$Produtividade
indv<- as.factor(rep(c(1:120),4) )

# Grupo 1 - resistente a ferrugem
# Grupo 2 - suscetivel a ferrugem
# Grupo 3 - São novas progênies
grupo<-as.factor(rep(c(rep(1,92),rep(2,8),rep(3,16),rep(2,4)),4))

# Organizando dados em uma Dataframe:
dados.f<-data.frame(indv,grupo,cult,ano,bloco,prod)

# Médias e Desvios padrão de produtividade para cada ano por cultivar
mean.cult<-tapply(prod,list(cult,ano),mean)
var.cult<-tapply(prod,list(cult,ano),var)
dp.cult<-sqrt(var.cult)
x<-cbind("Média 2010"=mean.cult[,1],"DP 2010"=dp.cult[,1],"Média 2011"=mean.cult[,2],
"DP 2011"=dp.cult[,2],"Média 2012"=mean.cult[,3],"DP 2012"=dp.cult[,3],
"Média 2013"=mean.cult[,4],"DP 2014"=dp.cult[,4])
# Anual:
var<-tapply(prod,ano,var)
dp<-var1/2
tapply(prod,cult,summary)
mediaano<-tapply(prod,ano,summary)

# Produtividade por cultivar:
summary.cult<-tapply(prod,cult,summary)
```

```

medias<-tapply(prod,cult,mean)
variancia<-tapply(prod,cult,var)
desvio<-(variancia)1/2

# Cultivares resistentes 1-23, cultivares controle suscetíveis 24,25 e 30, progênes ex-
perimentais 26-29
# Grupo 1 - resistente a ferrugem
# Grupo 2 - suscetível a ferrugem
# Grupo 3 - São novas progênes
tapply(prod,grupo,mean)
tapply(prod,grupo,var)

# Gráficos

# Boxplot da produtividade Cultivares
boxplot(prod cult,col = "bisque")

# Bloxplot da produtividade bloco
boxplot(prod bloco)

# Box-plot da produtividade ano
boxplot(prod ano)

# Boxplot produtividade grupo
boxplot(prod grupo)

# Gráfico de perfis médios de produtividade por cultivar
G1<- groupedData(prod ano.n|cult, order.groups=F,data=dados.f)
plot(G1, display="CCS",aspect=2,key=F,xlab=("Tempo(ano)"),
ylab=("Produtividade(Sc/ha)"),xaxp=c(0,2,2014))
# Gráfico de perfis individuais das produtividades
G2<-groupedData(prod ano.n|cult/indv, outer= cult,data=dados.f,order.groups=F)
plot(G2,outer=T,display=1,xlab=("Tempo(ano)"),ylab=("Produtividade(Sc/ha)"))

# Produtividade dos 30 tipos de cultivares nos 4 anos.
G3<- groupedData(prod ano|indv, order.groups=F,data=dados.f)
plot(G3, display= "CCS",aspect=2,key=T,xlab=("Produtividade(Sc/ha)"),
ylab=("indivíduo"),xaxp=c(0,2,2014))

# intervalos de confiança para intercepto e anos.
list<-lmList(prod ano|cult,data=G3)
plot(intervals(list))

# Matriz de correlação e covariância
Mprod<-cbind ( prod[361:480], prod[241:360], prod[121:240], prod[1:120] )
Mvarcov<- cov(Mprod)
Mvarcov[2,1]<-cor(Mprod)[2,1]
Mvarcov[3,1]<-cor(Mprod)[3,1]
Mvarcov[4,1]<-cor(Mprod)[4,1]
Mvarcov[3,2]<-cor(Mprod)[3,2]

```



```
Mvarcov[4,2]<-cor(Mprod)[4,2]
Mvarcov[4,3]<-cor(Mprod)[4,3]

# Modelo M1 por REML
M1<-lme(prod ~cult+ano+bloco+(cult*ano),
random= ~ano|indv, method="REML",
data=dados.f,
control=lmeControl(opt="optim"))
summary(M1)
anova(M1,test="wald")

# Modelo M1.1 por ML
M1.1<-update(M1,method="ML")
anova(M1.1,test="wald")

# Modelo M2 por REML
M2<-lme(prod ~cult+ano+bloco,
random= ~ano|indv, method="REML",
data=dados.f)
summary(M2)
anova(M2, test="wald")

# Modelo M2.1 por ML
M2.1<-update(M2,method="ML")

# Modelo M3 sem efeito fixo de Bloco
M3<- lme(prod ~cult+ano,
random= ~ano|indv, method="REML",
data=dados.f)
summary(M3)
anova(M3,test="wald")

# Modelo M3.1 por ML
M3.1<-update(M3,method="ML")

# Comparando modelos
anova(M2.1,M3.1)
anova(M1.1,M2.1,M3.1)

# Estruturas de covariância G

# Não Estruturada (UN)
M1.2<-lme(prod ~cult+ano+bloco+(ano*cult),
random=pdSymm( ~ano), method="ML",
data=G3)
summary(M1.2)
anova(M1.2, test="wald")
```

```
# Diagonal
M1.3<-update(M1.2, random=pdDiag( ano))
summary(M1.3)
anova(M1.3, test="wald")
# Comparando modelos
anova(M1.3, M1.2 )

# Matriz de variância covariância efeitos aleatórios G
getVarCov(M1.2,type="random.effects")
# Estruturas de covariância Ri

# Não Estruturada. M1.4<-lme(prod  cult+ano+bloco+(ano*cult), random=pdDiag( ano),
method="REML", corr=corSymm(form= 1|indv), data=G3)

# simetria composta M1.5<-lme(prod  cult+ano+bloco+(ano*cult), ran-
dom=pdDiag( ano), method="REML", corr=corCompSymm(form= 1|indv), data=G3)
# Correlação Ar1 M1.6<-lme(prod  cult+ano+bloco+(ano*cult), random=pdDiag( ano),
method="REML", corr=corAR1(form= 1|indv), data=G3)
# Tukey comparar cultivares mais produtivas:8,21,7, 17

k1<- c(rep(0,7),1,rep(0,12),-1,rep(0,9)) 8 com 21
k2<- c(rep(0,6),1,-1,rep(0,22)) 8 com 7
k3<- c(rep(0,6),1,0,rep(0,12),-1,rep(0,9)) 21 com 7
k<-rbind(k1,k2,k3)
summary(glht(M1.4, linfct=mcp(cult=k)))

# Diagnóstico do modelo M1.4

# Teste de normalidade Shapiro Wilk
shapiro.test(res.M1.4)

# Teste de Kolmogorov-Smirnov
ks.test(res.M1.4, "pnorm",mean(res.M1.4),sd(res.M1.4))

plot(fitted(M1.4),prod, xlab="Valores preditos - M1.4", ylab="Produtividade em sacas/ha")
abline(0,1)

plot(fitted(M1.4),residuals(M1.4, type="p"), ylim=c(-4,4), xlab="Valores preditos - M1.4",
ylab="Resíduos Condicionais Padronizados")
abline(h=c(-3,0,3), lty= c(2,1,2))

qqnorm(res.M1.4,xlab="Quantis Normais",ylab="Resíduos Condicionais Padronizados -
M1.4", main=)
qqline(res.M1.4,col="red")

hist(resid(M1.4, type="p"),main=,xlab="Resíduos Condicionais Padronizados - M1.4",
ylab="Frequência")

plot(M1.4, resid(.)  fitted(.) | cult, abline = 0)
```