Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Seleção de modelos lineares mistos utilizando critérios de informação

Tatiana Kazue Yamanouchi

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2017

Tatiana Kazue Yamanouchi
Licenciada em Matemática

Seleção de modelos lineares mistos utilizando critérios de informação

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **CÉSAR GONÇALVES DE LIMA**

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Yamanouchi, Tatiana Kazue
Seleção de modelos lineares mistos utilizando critérios de informação / Tatiana Kazue Yamanouchi. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2017 .
57 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Modelos mistos 2. Seleção de modelos 3. Simulação 4. Critério de informação . I. Título.

DEDICATÓRIA

Aos meus pais Massamori e Helena,
às minhas irmãs Tiemy e Satie
e ao meu namorado Eliton,
dedico.

AGRADECIMENTOS

Agradeço, primeiramente a Deus pela vida.

Agradeço meus pais Massamori e Helena, por sempre me apoiar em todos os momentos da minha vida, e as minhas irmas Tiemy e Satie, por sempre me incentivarem e pela amizade.

Ao meu namorado, Eliton Moro, por sempre estar do meu lado, por me ajudar sempre, pelo companheirismo, amizade e amor.

Ao meu orientador, professor Dr. César Gonçalves de Lima, pelos ensinamentos, paciência e amizade.

Aos docentes e funcionários do Departamento de Ciências Exatas, pela atenção e pelos ensinamentos. Em especial, agradeço ao professor Dr. Cristian Marcelo Villegas Lobos, por todos os ensinamentos, principalmente no uso do software R e pela disponibilidade em me ajudar sempre.

À família do Eliton, em especial, Paula, Junior e Elton, por me acolherem como membro da família, pelo carinho e amor.

À minha madrinha, Hiromi, por todo apoio e ajuda.

À minha família, em especial ao tio Hiroki e tia Youko (*in memoriam*) aos meus avós, Hiroko e Hidenori (*in memoriam*), por terem me acolhido.

À tia Tô e tio Paulo por terem me acolhido em São Paulo.

Aos meus amigos da graduação da Unesp Rio Claro e aos colegas do mestrado, especialmente a Michele Penso pelos 7 anos de amizade e convívio. Agradeço também ao Júlio pela amizade durante o mestrado.

Ao CNPq pelo apoio financeiro.

À todos que, diretamente ou indiretamente, contribuíram para a realização deste trabalho.

EPÍGRAFE

"Descobrir consiste em olhar para o que todo mundo está vendo e pensar uma coisa diferente".
(Roger Von Oech)

SUMÁRIO

RESUMO

Seleção de modelos lineares mistos utilizando critérios de informação

O modelo misto é comumente utilizado em dados de medidas repetidas devido a sua flexibilidade de incorporar no modelo a correlação existente entre as observações medidas no mesmo indivíduo e a heterogeneidade de variâncias das observações feitas ao longo do tempo. Este modelo é composto de efeitos fixos, efeitos aleatórios e o erro aleatório e com isso na seleção do modelo misto muitas vezes é necessário selecionar os melhores componentes do modelo misto de tal forma que represente bem os dados. Os critérios de informação são ferramentas muito utilizadas na seleção de modelos, mas não há muitos estudos que indiquem como os critérios de informação se desempenham na seleção dos efeitos fixos, efeitos aleatórios e da estrutura de covariância que compõe o erro aleatório. Diante disso, neste trabalho realizou-se um estudo de simulação para avaliar o desempenho dos critérios de informação AIC, BIC e KIC na seleção dos componentes do modelo misto, medido pela taxa TP (Taxa de verdadeiro positivo). De modo geral, os critérios de informação se desempenharam bem, ou seja, tiveram altos valores de taxa TP em situações em que o tamanho da amostra é maior. Na seleção de efeitos fixos e na seleção da estrutura de covariância, em quase todas as situações, o critério BIC teve um desempenho melhor em relação aos critérios AIC e KIC. Na seleção de efeitos aleatórios nenhum critério teve um bom desempenho, exceto na seleção de efeitos aleatórios em que considera a estrutura de simetria composta, situação em que BIC teve o melhor desempenho.

Palavras-chave: Modelos mistos; Seleção de modelos; Critério de informação; Simulação

ABSTRACT

Mixed linear model selection using information criterion

The mixed model is commonly used in data of repeated measurements because of its flexibility to incorporate in the model the correlation existing between the observations measured in the same individual and the heterogeneity of variances of observations made over time. This model is composed of fixed effects, random effects and random error and with this in the selection of the mixed model it is often necessary to select the best components of the mixed model in such a way that it represents the data well. Information criteria are tools widely used in model selection, but there are not many studies that indicate how information criteria play out in the selection of fixed effects, random effects, and the covariance structure that makes up the random error. In this work, a simulation study was performed to evaluate the performance of the AIC, BIC and KIC information criteria in the selection of the components of the mixed model, measured by the TP (True positive Rate). In general, the information criteria performed well, that is, they had high TP rate in situations where the sample size is larger. In the selection of fixed effects and in the selection of the covariance structure, in almost all situations, the BIC criterion had a better performance in relation to the AIC and KIC criteria. In the selection of random effects no criterion had a good performance, except in the selection of Random effects in which it considers the compound symmetric structure, situation in which BIC had the best performance.

Keywords: Mixed models; Model selection; Information criterion; Simulation

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MV	Método da Máxima Verossimilhança
MVR	Método da Máxima Verossimilhança Restrita
VC	Componente de Variância
UN	Não Estruturada
ARH(1)	Auto - Regressiva com Heterogeneidade de variância
AR(1)	Auto - Regressiva de ordem 1
CS	Simetria Composta
TP	Taxa verdeiro positivo (*True Positive*)

1 INTRODUÇÃO

Estudos envolvendo dados com medidas repetidas estão presentes em vários campos de pesquisa, como a medicina, ciências biológicas, economia, agropecuária, etc. A estrutura desses dados tem como característica a presença de duas ou mais observações da variável resposta em cada unidade amostral sob investigação. O estudo longitudinal envolve medidas repetidas ao longo do tempo, ao passo que o estudo de medidas repetidas além de envolver essa característica, englobam estudos de experimentos como *split-plot* e com intercâmbio (*cross-over*).

Alguns interesses no estudo de dados longitudinais são de avaliar as mudanças globais ou individuais ao longo do tempo, sendo comum admitir uma correlação não nula entre as observações feitas em instantes distintos, já que as medidas são feitas em uma mesma unidade experimental e também uma heterogeneidade de variâncias das medidas feitas em diferentes ocasiões.

Dentre as diversas abordagens usadas na análise estatística desse tipo de dados, o modelo linear misto (Verbeke & Molenberghs, 2000) tem sido utilizado com êxito, pois consegue explicar bem o comportamento da média da variável resposta por meio dos parâmetros de efeito fixo, a variabilidade inerente aos indivíduos (entre-indivíduos) por meio dos parâmetros de efeito aleatório e da variabilidade do erro aleatório (intra-indivíduos).

Na busca por um modelo que represente bem o fenômeno em estudo podem ser encontrados vários modelos plausíveis. Neste caso, existe a necessidade de se utilizar algum método ou alguma estatística que auxilie nesta escolha. Na seleção do modelo linear misto mais adequado são comumente usados o teste da razão de verossimilhança e os critérios de informação, sendo o critério de informação de Akaike - AIC (Akaike, 1974) e o critério de informação Bayesiano ou Schwarz - BIC (Schwarz, 1978). O critério de informação de Kullback (Cavanaugh, 1999) também pode ser usado nesta escolha.

O interesse na seleção dos modelos mistos é de encontrar um modelo adequado que explique bem o comportamento da resposta média ao longo do tempo e a estrutura de variâncias e covariâncias entre as medidas repetidas, ou seja, é necessário fazer seleção dos efeitos fixos e aleatórios.

Neste trabalho avaliou-se a performance dos critérios de informação, AIC, BIC e de Kullback - KIC (Cavanaugh,1999) na seleção dos modelos mistos (efeitos fixos, aleatórios e a estrutura de covariâncias). Para avaliar as performances dos critérios de informação utilizou-se a taxa de verdadeiro positivo (TP) em um estudo de simulação considerando diferentes cenários, a partir de um conjunto de dados de pesos semanais de frangos num período de 7 semanas, utilizados por Lima (1988) e por Barbosa (2009). Na simulação utilizou-se o *software* R.

No capítulo 2 apresenta-se uma revisão de modelos lineares mistos envolvendo os métodos de estimação, estruturas de covariâncias e principais processos de inferência. É apresentado também a seleção de modelos, utilização dos *softwares* R e SAS nos ajustes do modelo linear misto e a avaliação do desempenho dos critérios de informação.

No capítulo 3 apresenta-se o material utilizado e a descrição dos cenários usados no estudo de simulação.

No capítulo 4 apresenta-se alguns resultados e discussões.

No capítulo 5 apresenta-se as conclusões obtidas no trabalho.

2 REVISÃO DE LITERATURA

Nesta seção serão apresentados uma revisão geral sobre os modelos mistos, incluindo a especificação do modelo, métodos de estimação (Método da Máxima Verossimilhança (MV) e Método da Máxima Verossimilhança Restrita (MVR) e também, uma revisão sobre seleção de modelos e critérios de informações.

2.1 Definições e exemplos

Em pesquisas científicas é comum encontrar variáveis que são correlacionadas. Este comportamento está presente em várias estruturas de dados, especialmente nos dados com medidas repetidas.

De acordo com Singer, Nobre e Rocha, (2015), o estudo de planejamento longitudinal (conhecido também como *coorte* na área de bioestatística e *painel* na área de sociologia, administração ou economia) envolve a realização de duas ou mais observações da variável resposta em cada unidade amostral sob investigação.

Este tipo de estudos é um caso particular de um estudo de medidas repetidas, que engloba experimentos como *split-plot* e com intercâmbio (*cross-over*).

O planejamento longitudinal é dito balanceado com relação ao tempo em situações em que as observações em todas as unidades amostrais são feitas nos mesmos instantes, quer sejam igualmente espaçados ou não. Nos casos em que houver observações de diferentes unidades amostrais em diferentes instantes, o planejamento longitudinal é dito desbalanceado com relação ao tempo.

Neste tipo de estudo, o principal interesse consiste em analisar características populacionais e individuais ao longo do "tempo"e neste caso, a utilização dos modelos mistos se torna uma ferramenta importante.

O termo "tempo"não necessariamente se refere ao tempo (dias, anos, minutos, etc.), mas se refere a uma escala ao longo da qual são feitas as medidas repetidas, conforme definido por Singer, Nobre e Rocha (2015). Neste trabalho, os autores apresentaram um estudo realizado na Escola de Educação Física e Esporte da Universidade de São Paulo, cujo objetivo era avaliar o efeito de um programa de treinamento realizado por idosos considerando os gêneros (masculino e feminino). A atividade que teve foco nesse estudo foi de medir o tempo gasto pelos idosos em calçar as meias antes do treinamento e após o treinamento.

Este exemplo é caracterizado como um estudo longitudinal, pois existem duas medidas feitas no mesmo indivíduo (idoso): o tempo gasto em calçar a meia antes do treinamento e o tempo gasto depois do treinamento. Nesse estudo pode-se avaliar o efeito do fator gênero, que tem dois níveis: masculino e feminino e existem dados omissos, ou seja, há indivíduos que não tiveram marcação antes ou depois do treinamento aplicado, o que permite dizer que os dados são desbalanceados em relação ao tempo.

2.2 Modelo Linear Misto

De acordo com Verbeke & Molenberghs (2000), um modelo linear misto geral pode ser formulado em dois estágios:

- Estágio 1

  Seja Y_{ij}, a variável resposta para cada indivíduo i, medida no tempo t_{ij}, com $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ e seja $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ o vetor de variáveis respostas para cada indivíduo i.

  O modelo do primeiro estágio tem como princípio modelar a variável resposta para cada indivíduo i ao longo do tempo por meio de um modelo de regressão, do tipo

$$\mathbf{Y}_i = \mathbf{Z}_i\beta_i + \varepsilon_i \tag{2.1}$$

em que $\mathbf{Z}_i$ é uma matriz de dimensão $(n_i \times q)$ de variáveis regressoras conhecidas, β_i um vetor $(q \times 1)$ dos parâmetros de regressão e $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})'$ um vetor de erros de dimensão $(n_i \times 1)$. Assume-se que os ε_i, $i = 1, \ldots, m$ são independentes e seguem distribuição normal multivariada com vetor de média $\mathbf{0}$ e matriz de variâncias e covariâncias $\sigma^2\mathbf{I}_{n_i}$.

- Estágio 2

  O modelo do segundo estágio tem como objetivo explicar a variabilidade entre os indivíduos

$$\beta_i = \mathbf{K}_i\beta + \mathbf{b}_i \tag{2.2}$$

em que $\mathbf{K}_i$ é uma matriz de covariáveis conhecidas $(q \times p)$, β um vetor $(p \times 1)$ de parâmetros de regressão desconhecidos e assume-se que $\mathbf{b}_i$ é independente e segue distribuição normal multivariada $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{G})$.

Combinando os dois estágios segue o modelo

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i \quad i = 1, \ldots, m \tag{2.3}$$

em que $\mathbf{Y}_i$ (conhecido como perfil individual de resposta) é um vetor $(n_i \times 1)$, β é um vetor $(p \times 1)$ de parâmetros de efeitos fixos, $\mathbf{X}_i = \mathbf{K}_i\mathbf{Z}_i$ representa uma matriz $(n_i \times p)$ de especificação dos efeitos fixos (de posto completo), $\mathbf{b}_i$ é um vetor $(q \times 1)$ dos efeitos aleatórios, $\mathbf{Z}_i$ representa uma matriz $(n_i \times q)$ de especificação dos efeitos aleatórios (de posto completo) e ε_i é um vetor $(n_i \times 1)$ de erros aleatórios. Admite-se que $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{G})$, $\varepsilon_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i)$ e que $\mathbf{b}_i$ e ε_i são independentes. As matrizes $\mathbf{G}$, de dimensão $(q \times q)$ e $\mathbf{R}_i$, de dimensão $(n_i \times n_i)$, são matrizes simétricas positivas definidas sendo que a matriz $\mathbf{G}$ está relacionada com a variação entre os indivíduos, ao passo que a matriz de variância e covariância $\mathbf{R}_i$ está relacionada com a variação dentro do indivíduo, ou seja, ao longo do tempo.

Do modelo (2.3), segue que o modelo hierárquico $\mathbf{Y}_i|\mathbf{b}_i$ segue distribuição $N(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i)$. Dado que $\mathbf{Y}_i = \int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i$, resulta que o modelo marginal $\mathbf{Y}_i$ segue distribuição normal com média $\mathbf{X}_i\beta$ com matriz de covariância $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$.

Outra maneira de especificar o modelo é considerar o modelo com todos os indivíduos

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon \tag{2.4}$$

em que $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_m')'$ é um vetor de dimensão $(N \times 1)$ da variável resposta em que $N = \sum_{i=1}^{m} n_i$; a matriz $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_m')$ de dimensão $(N \times p)$ com $\mathbf{X}_i$ tendo dimensão $(n_i \times p)$, a matriz $\mathbf{Z} = \bigoplus_{i=1}^{m} \mathbf{Z}_i$ com dimensão $(N \times mq)$, em que $\bigoplus$ representa a soma direta em que cada $\mathbf{Z}_i$ tem dimensão $(n_i \times q)$, $\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2', \ldots, \mathbf{b}_m')'$ é um vetor de efeitos aleatórios de dimensão $(mq \times 1)$ e $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_m')'$ é um vetor de dimensão $(N \times 1)$. Sob esta formulação, tem-se que $\mathbf{b} \sim N_{mq}(\mathbf{0}, \Gamma)$, em que $\Gamma = \mathbf{I}_m \bigotimes_m \mathbf{G}$, com $\bigotimes$ representando o produto Kronecker e $\mathbf{b}$ é independente de $\varepsilon \sim N_N(\mathbf{0}, \mathbf{R})$, com $\mathbf{R} = \bigoplus_{i=1}^{m} \mathbf{R}_i$ e, consequentemente $\mathbf{y} \sim N_N(\mathbf{X}\beta, \mathbf{V})$, em que $\mathbf{V} = \mathbf{Z}\Gamma\mathbf{Z}' + \mathbf{R}$. (Singer, Nobre e Rocha, 2015).

2.3 Estruturas da Matriz de Covariância

De acordo com Singer, Nobre e Rocha (2015), as covariâncias entre as observações feitas no mesmo indivíduo podem ser modeladas indiretamente pelos efeitos aleatórios $\mathbf{b}_i$, que representam a variabilidade entre as unidades amostrais e diretamente por $\mathbf{R}_i$ que representa a variabilidade das observações intraunidades amostrais, ou seja, entre as observações realizadas nas diversas ocasiões. Vonesh & Chinchilli (1997) afirmam que a covariância entre as observações no mesmo indivíduo podem ser modeladas

como uma combinação de $\mathbf{b}_i$ e $\mathbf{R}_i$ permitindo a inclusão das três fontes de variação na estrutura de covariância: efeitos aleatórios, a variação devida a erros de medida na matriz $\mathbf{G}$ e a correlação serial que é representada na matriz $\mathbf{R}_i$.

As matrizes $\mathbf{G}$, $\mathbf{R}_i$ e $\mathbf{V}_i$ podem assumir diversas estruturas. Para exemplificar, vamos apresentar algumas estruturas, para o caso particular de $n_i = n = 3$ tempos:

1. Componente de Variância (ou diagonal)

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

   Esta estrutura assume variâncias iguais nos tempos e independência entre as observações repetidas. Neste caso, o número de parâmetros é igual a $r = 1$.

2. Não Estruturada (*Unstructured*)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

   Esta estrutura representa uma situação em que não se tem evidências de padrões sistemáticos das variâncias e das covariâncias. O número de parâmetros é de $r = \frac{n(n+1)}{2}$ em que n é o número de instantes de tempos. Neste caso, o número de parâmetros é $r = 6$.

3. Simetria Composta (*Compound Simmetry*)

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

   Essa estrutura possui $r = 2$ parâmetros admitindo variâncias homogêneas e covariâncias constantes entre as medidas repetidas. Esta estrutura é muito comum em ensaios em blocos casualizados quando se considera o efeito de blocos como aleatório e também em experimentos com parcelas subdivididas.

4. AR(1) (*first - order autoregressive structure*)
   Esta estrutura formaliza a ideia de decaimento das correlações entre as observações repetidas à medida que ficam mais distantes, como mostra a matriz de correlação Γ.

$$\Gamma = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

   No caso em que se assume variâncias iguais e combinando com a matriz de correlação, tem-se que a estrutura AR(1) segue a forma

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

   em que $-1 \leq \rho \leq 1$ e $\sigma > 0$.

   Neste caso, observe que o número de parâmetros é igual a $r = 2$ e esta estrutura pode ser utilizada somente em situações em que as observações repetidas são igualmente espaçadas.

5. ARH(1)

   Da mesma maneira que a estrutura AR(1), a estrutura ARH(1) também formaliza a ideia de decaimento das correlações entre as observações que ficam mais distantes, são utilizadas em situações em que as observações são igualmente espaçadas e o parâmetro ρ assume valores $-1 \leq \rho \leq 1$. A diferença entre as duas estruturas se trata da homogeneidade de variâncias assumida na estrutura AR(1) e heterogeneidade de variância assumida na estrutura ARH(1).

   Considere σ_j^2 e σ_k^2 variâncias nos tempos t_j e t_k, respectivamente, e σ_{jk} a covariância entre os dois tempos. Tem-se que $\rho = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$ ou $\sigma_{jk} = \rho \sigma_j \sigma_k$ e portanto, segue que a matriz de covariância é dada por

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

   Neste caso, a estrutura da matriz de covariâncias tem $r = 4$ parâmetros.

6. Markov

   O modelo de correlação de Markov é utilizado quando as observações repetidas ao longo do tempo não são necessariamente igualmente espaçadas. A matriz de correlação é dada por

$$\Gamma = \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} \\ \rho^{d_{12}} & 1 & \rho^{d_{23}} \\ \rho^{d_{13}} & \rho^{d_{23}} & 1 \end{bmatrix}$$

   em que $d_{jk} = |t_j - t_k|$.

   Com a suposição de variâncias iguais e combinando com a matriz de correlação, tem-se que a matriz de covariâncias de Markov é dada por

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} \\ \rho^{d_{12}} & 1 & \rho^{d_{23}} \\ \rho^{d_{13}} & \rho^{d_{23}} & 1 \end{bmatrix}$$

7. Toeplitz

   A estrutura de Toeplitz provém de um processo de médias móveis de ordem t

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

   Neste caso, $t = 2$ e o número de parâmetros $r = 3$.

   Outras estruturas de matrizes de covariância podem ser encontradas em Pinheiro e Bates (2000), SAS/STAT (2008) dentre outros.

2.4 Estimação do modelo marginal

A variável aleatória $\mathbf{Y}_i$, apresentada na equação (2.3), segue distribuição normal multivariada com vetor de médias $\mathbf{X}_i\beta$ e matriz de covariâncias $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. Considere θ_g o vetor associado aos parâmetros da matriz $\mathbf{G}$, θ_r o vetor de parâmetros de $\mathbf{R}_i$, $\alpha = (\theta_g, \theta_r)'$ e finalmente $\theta = (\alpha, \beta)'$, em que α é o vetor de parâmetros das matrizes $\mathbf{G}$ e $\mathbf{R}_i$ e β é o vetor de parâmetros de regressão.

Dentre os métodos de estimação dos parâmetros do modelo (2.3) destacam-se os métodos de Máxima Verossimilhança (MV) e Máxima Verossimilhança Restrita (MVR), que serão apresentados com detalhes a seguir. Estes métodos de estimação podem ser encontrados em West, Welch & Gatecki (2006), Singer, Nobre e Rocha (2015), dentre outros.

2.4.1 Método da Máxima Verossimilhança (MV)

A estimação de máxima verossimilhança consiste na maximização da função de verossimilhança, que neste caso está baseada na distribuição do modelo marginal. Para cada indivíduo i, $\mathbf{Y}_i$ segue distribuição normal multivariada com função densidade de probabilidade definida por:

$$f(\mathbf{y}_i|\theta) = (2\pi)^{-\frac{n_i}{2}}|\mathbf{V}_i|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)\right\} \tag{2.5}$$

em que $|\mathbf{V}_i|$ representa o determinante da matriz $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$.

A função de verossimilhança, $L(\theta|\mathbf{y}_i)$, é dado por:

$$L(\theta|\mathbf{y}_i) = \prod_{i=1}^{m}(2\pi)^{-\frac{n_i}{2}}|\mathbf{V}_i|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)\right\} \tag{2.6}$$

e o logaritmo da função de verossimilhança, $l(\theta|\mathbf{y_i})$

$$l(\theta|\mathbf{y_i}) = ln\{L(\theta|\mathbf{y}_i))\} = -\frac{N}{2}ln(2\pi) - \frac{1}{2}\sum_{i=1}^{m}ln(|\mathbf{V}_i|) - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta) \tag{2.7}$$

em que $N = \sum_{i=1}^{m} n_i$ e ln representa o logaritmo na base e.

Assumindo que a matriz $\mathbf{V}_i$ é conhecida, derivando a equação (2.7) em relação a β e igualando a um vetor de zeros tem-se que

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i\mathbf{V}_i^{-1}\mathbf{y}_i \tag{2.8}$$

que é chamado de BLUE (Melhor Estimador Linear Não Viesado) de β.

Dado que α é desconhecido, encontra-se uma estimativa $\hat{\alpha}$ de α. Para tal, substitui-se $\hat{\beta}(\alpha)$, obtido na equação (2.8), na função de log verossimilhança apresentada na equação (2.7), obtendo a função log verossimilhança perfilada $l(\hat{\beta}(\alpha), \alpha|\mathbf{y}_i)$. Derivando a função log verossimilhança perfilada em relação a α e igualando a um vetor de zeros, obtêm-se o estimador de máxima verossimilhança de α e substituindo o estimador de α na equação (2.8), encontra-se o estimador de máxima verossimilhança de β.

Como não existem formas fechadas dos estimadores, as estimativas dos parâmetros são obtidas por métodos iterativos como: Algoritmo de Newton - Raphson, Algoritmo EM (*Expectation - maximization*) e Algoritmo Escore de Fisher. Para mais detalhes consulte Gilmour, Thompson, Cullis(1995), Laird & Ware (1982), dentre outros.

2.4.2 Método da Máxima Verossimilhança Restrita (MVR)

Segundo West, Welch & Gatecki (2006), a estimação do vetor de parâmetros das matrizes de covariâncias de $\mathbf{G}$ e $\mathbf{R}_i$, α, por máxima verosimilhança produz estimativas viesadas, pois não leva em consideração a perda de graus de liberdade que resulta da estimação dos parâmetros de efeito fixo β. Uma alternativa para corrigir o viés produzido pela estimativa dos parâmetros de covariâncias consiste no uso do método da máxima verossimilhança restrita (MVR), em que as estimativas dos componentes de variância tendem a ser menos viesadas que as estimativas produzidas pelo MV.

O MVR foi proposto por Patterson & Thompson (1971) para estimar componentes de variância e consiste em maximizar a verossimilhança de uma transformação linear ortogonal $\mathbf{y}^* = \mathbf{U}'\mathbf{y}$, em que $\mathbf{U}$ é uma matriz de posto completo, de dimensão $(N \times N-p)$, com colunas ortogonais às colunas da matriz $\mathbf{X}$ e é tal que $E(\mathbf{y}^*) = \mathbf{0}$. A verossimilhança obtida por meio dessa transformação não depende do vetor de parâmetros β nem da escolha do matriz $\mathbf{U}$. Em geral, utiliza-se $\mathbf{U} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, que é a matriz de projeção que gera os resíduos do ajuste obtidos por mínimos quadrados ordinários (Singer, Nobre e Rocha, 2015).

Assim, $\mathbf{y}^* \sim N(\mathbf{0}, \mathbf{U}'\mathbf{V}\mathbf{U})$ e a função log-verossimilhança restrita, $l_R(\theta)$, é

$$l_R(\theta) = -\frac{1}{2}ln|\mathbf{V}| - \frac{1}{2}ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{N-p}{2}(\mathbf{y}-\mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta) - \frac{N-p}{2}ln|2\pi| \tag{2.9}$$

que também pode ser escrita como:

$$l_R(\theta) = -\frac{1}{2}ln(2\pi)\sum_{i=1}^{m} n_i - \frac{1}{2}\sum_{i=1}^{m} ln|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\hat{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}) - \frac{1}{2}ln|\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i| \tag{2.10}$$

Logo, a maximização da equação (2.9) ou (2.10), produz as estimativas de MVR de θ.

Outros detalhes sobre estes métodos de estimação, MV e MVR, pode ser encontradas em Casella & Berger (2002), Harville (1977), Demidenko (2013), dentre outros.

2.5 Estimação de efeito fixo e predição de efeitos aleatórios

As equações de modelos mistos (Henderson, 1953), baseadas na maximização da função de verossimilhança conjunta de $\mathbf{y}$ e $\mathbf{b}$, permitem obter as estimativas dos parâmetros de efeito fixo β e a predição dos efeitos aleatórios $\mathbf{b}$, simultaneamente.

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\Gamma^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{2.11}$$

Assim, pode-se mostrar que o estimador $\hat{\beta}$ é dado por

$$\hat{\beta}(\alpha) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right) \tag{2.12}$$

ou, alternativamente por

$$\hat{\beta}(\alpha) = \sum_{i=1}^{m}\left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i\right) \tag{2.13}$$

e a predição de $\mathbf{b}$ é dada por

$$\widehat{\mathbf{b}} = \Gamma\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{2.14}$$

ou, alternativamente por

$$\widehat{\mathbf{b}} = \mathbf{G}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}) \tag{2.15}$$

em que $\hat{\beta}$ é o BLUE (Melhor Estimador Linear Não Viesado) de β e $\hat{\mathbf{b}}$ é o BLUP (Melhor Preditor Linear Não Viesado) de $\mathbf{b}$. Os estimadores $\hat{\beta}$ e $\hat{\mathbf{b}}$ dependem dos parâmetros de covariância α, sendo assim, calcula-se o BLUE e o BLUP com base no estimador $\hat{\alpha}$ de α. Estes são chamados de EBLUE (Melhor Estimador Linear Não Viesado Empírico) de β e $\widehat{\mathbf{b}}$ é EBLUP (Melhor Preditor Linear Não Viesado Empírico) de $\mathbf{b}$.

Singer, Nobre e Rocha (2015) comentam que existem diversas vantagens na obtenção do BLUE e do BLUP por meio das equações de modelos mistos, sendo útil por exemplo, nas técnicas de diagnóstico.

2.6 Inferência

Um dos objetivos da análise de dados com medidas repetidas consiste em fazer inferências sobre os parâmetros do modelo, quer sejam os parâmetros de efeito fixo ou mesmo os parâmetros de covariâncias. Geralmente as inferências sobre os parâmetros de um modelo linear misto se baseiam em distribuições aproximadas para os estimadores de MV ou de MVR devido a seus resultados assintóticos (Pinheiro e Bates, 2000).

Pinheiro (1994) mostrou que sob certas condições de regularidade, ou seja, o espaço paramétrico deve ter dimensão finita, ser fechado e compacto e que o verdadeiro valor do parâmetro esteja em seu interior (Singer, Nobre e Rocha, 2015), tanto os estimadores obtidos por MV quanto por MVR do modelo (2.3), são consistentes e assintoticamente normais. Esses resultados assintóticos auxiliam na construção de testes de hipóteses sobre os parâmetros de covariância α e de efeitos fixos β.

A seguir serão apresentados alguns testes para os efeitos fixos.

1. Teste Wald aproximado para a hipótese:

   $$H_0 : \mathbf{L}\beta = \mathbf{0} \; vs \; H_a : \mathbf{L}\beta \neq \mathbf{0}$$

   considera que a estatística

   $$\mathbf{W} = (\hat{\beta} - \beta)'\mathbf{L}'(\mathbf{L}(\sum_i \mathbf{X}_i'\mathbf{V}_i'\mathbf{X}_i')^{-1}\mathbf{L}')\mathbf{L}(\hat{\beta} - \beta)$$

   segue distribuição χ^2 com os graus de liberdade igual ao $posto(\mathbf{L})$, em que $\mathbf{L}$ é uma matriz de constantes e de posto-linha completo e $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$.

2. Teste t

   O teste t é utilizado para testar hipóteses como

   $$H_0 : \beta_j = 0 \; vs \; H_a : \beta_j \neq 0$$

   com $j = 1, \ldots, p$. A estatística t definida por

   $$t = \frac{\beta_j}{\sqrt{var(\hat{\beta})}}$$

   segue distribuição t aproximada, com número de graus de liberdade estimado, por exemplo, pelo método de Satterthwaite (1946), dentre outros.

3. Teste Wald F

   Considere a mesma especificação do teste de hipótese do Teste Wald aproximado. A estatística

   $$F = \frac{\mathbf{W}}{posto(\mathbf{L})}$$

   em que $\mathbf{W}$ é a estatística do teste Wald aproximado, segue distribuição aproximada $F(\upsilon_1, \upsilon_2)$, em que $\upsilon_1 = posto(\mathbf{L})$ e υ_2 pode ser estimado utilizando métodos como os propostos por Satterthwaite (1946) e Kenward & Roger (1997). Note que alguns desses métodos substituem $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$ pela sua estimativa $\hat{\mathbf{V}}_i$.

4. Teste da razão de verossimilhanças (TRV)

   O teste da razão de verossimilhanças é usado para comparar modelos aninhados com diferentes estruturas de média.

$$H_0 : \beta \in \Theta_{\beta,0} \; vs \; H_a : \beta \in \Theta_\beta$$

   para algum subespaço $\Theta_{\beta,0}$ do espaço paramétrico Θ_β de efeitos fixos β.

   Associando o modelo aninhado à hipótese nula e o modelo referência à hipótese alternativa, tem-se que a estatística do teste de razão de verossimilhança é dada por

$$-2ln\left(\frac{L_{aninhado}}{L_{referência}}\right) = 2ln(L_{referência}) - 2ln(L_{aninhado}) \tag{2.16}$$

   que sob certas condições de regularidade (Singer, Nobre e Rocha, 2015), segue, assintoticamente, a distribuição χ^2_v, em que v é a diferença entre os números de parâmetros dos modelos de referências e aninhado, $L_{aninhado}$ é o valor da função de verossimilhança MV, baseada no subespaço $\Theta_{\beta,0}$ e $L_{referência}$ é o valor da função de verossimilhança MV, baseada no espaço Θ_β.

Quando o interesse do teste consiste em comparar dois modelos aninhados que diferem apenas na especificação da parte fixa do modelo, deve-se considerar apenas o ajuste dos modelos considerando o método da máxima verossimilhança (Pinheiro e Bates, 2000). No entanto estes mesmos autores aconselham a não utilizar o TRV para a especificação da parte fixa do modelo, mas sim o teste F e o teste t.

Verbeke & Molenberghs (2000) também recomendam o uso do teste t ou do teste F para testar hipóteses sobre os parâmetros de efeito fixo ao invés do teste Wald aproximado.

Agora serão apresentados alguns testes para os componentes de variância.

1. Teste Wald aproximado

   Análogo ao teste Wald aproximado apresentado nos testes de efeitos fixos, o teste para a hipótese

$$H_0 : \mathbf{L}\alpha = \mathbf{0} \; vs \; H_a : \mathbf{L}\alpha \neq \mathbf{0}$$

   tem-se que, sob certas condições de regularidade, a estatística $\mathbf{W}$ segue distribuição χ^2 com os graus de liberdade igual ao $posto(\mathbf{L})$, em que $\mathbf{L}$ é uma matriz qualquer de posto-linha completo e α é o vetor de parâmetros da matriz de covariância.

2. Teste da razão de verossimilhança

   De modo análogo ao teste da razão de verossimilhanças apresentado nos testes de hipóteses para efeitos fixos, o teste para a hipótese

$$H_0 : \alpha \in \Theta_{\alpha,0} \; vs \; H_a : \alpha \in \Theta_\alpha$$

   para algum subespaço $\Theta_{\alpha,0}$ do espaço paramétrico Θ_α de componentes de variância α, tem-se que a estatística do teste da razão de verossimilhanças, sob certas condições de regularidade e sob H_0, segue assintoticamente a distribuição χ^2_v, em que v é a diferença entre os números de parâmetros dos modelos de referência e aninhado.

   Adicionalmente ao teste da razão de verossimilhanças para efeitos fixos, o teste para componentes de variância também é válido utilizando o método de estimação MVR. Neste caso, $L_{aninhado}$ e $L_{referência}$ são funções de verossimilhanças restritas, cujas estimativas são obtidas maximizando as

funções pelo método da máxima verossimilhança restrita. Pinheiro e Bates (2000) ressaltam que ao utilizar o método MVR, os dois modelos a serem comparados (aninhado e referência), devem ser ambos ajustados pelo método MVR e que a especificação da parte fixa do modelo deve ser a mesma em ambos os modelos.

Em muitos casos, são consideradas hipóteses do tipo

$$H_0 : \sigma_a^2 = 0 \; vs \; H_a : \sigma_a^2 \in (0, \infty)$$

com σ_a^2 sendo um dos componentes de α, que se encontra na fronteira do espaço paramétrico. Nestes casos, tanto o teste de Wald aproximado quanto o teste da razão de verossimilhanças, apresentam problemas quanto às condições de regularidade. Nessas condições, Self & Liang (1987) mostraram que a estatística do teste da razão de verossimilhanças pode ser aproximada por uma mistura de distribuições qui-quadrado.

2.7 Seleção de modelos

Segundo Barbosa (2009), a seleção do melhor modelo misto consiste na escolha da melhor estrutura para a parte fixa do modelo e da melhor estrutura de covariâncias. Para realizar esta seleção pode-se utilizar testes de hipóteses para os parâmetros de modelos aninhados, como apresentado na seção 2.6, ou utilizar os critérios de informação, que podem ser utilizados mesmo quando os modelos não são aninhados.

Emiliano (2013) afirma que ao estudar um fenômeno, este pode ser explicado por modelos, mas dificilmente consegue-se obter informações completas sobre o fenômeno. Ao tentar explicar tal fenômeno por um modelo probabilístico, podem existir vários modelos plausíveis que expliquem bem o mesmo fenômeno, e se busca escolher o modelo que mais se aproxima da realidade, ou seja, aquele modelo em que houve menor perda de informação. Neste contexto, o objetivo da seleção consiste em encontrar um modelo que minimize essa perda de informação.

Para quantificar essa perda de informações existem algumas medidas propostas na literatura. As estatísticas apresentadas a seguir foram citadas em Emiliano (2009), com exceção da estatística divergência simétrica de Kullback.

1. Estatística de χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{g_i^2}{f_i^2} - 1 \tag{2.17}$$

em que f_i e g_i são funções densidade de probabilidade quaisquer.

2. Informação Generalizada

$$I_\lambda(g; f) = \frac{1}{\lambda} \int_{-\infty}^{+\infty} \left[\left(\frac{g(x)}{f(x)} \right)^\lambda - 1 \right] g(x) dx \tag{2.18}$$

em que $\lambda > 0$ e f e g são funções densidade de probabilidade quaisquer.

3. Informação de Kullback-Leibler

$$I(g; f) = E_g \left[ln \left(\frac{g(X)}{f(X)} \right) \right] = \int_{-\infty}^{+\infty} g(x) ln \left(\frac{g(x)}{f(x|\theta)} \right) dx = \int_{-\infty}^{+\infty} g(x) ln(g(x)) dx - \int_{-\infty}^{+\infty} g(x) ln(f(x|\theta)) dx \tag{2.19}$$

em que g e f são funções densidade de probabilidade quaisquer, θ é o vetor de parâmetros pertencente ao espaço paramétrico Θ e $x \in R$.

4. Divergência Simétrica de Kullback

   Assim como $I(g;f)$ definido na equação (2.19), pode-se definir a informação de Kullback-Leibler entre f e g por $I(f;g)$. Dessa forma, define-se a divergência simétrica de Kullback por

$$J(g;f) = I(g;f) + I(f;g) = E_g\left[ln\left(\frac{g(X)}{f(X)}\right)\right] = E_f\left[ln\left(\frac{f(X)}{g(X)}\right)\right] \tag{2.20}$$

   em que g e f são funções densidade de probabilidade quaisquer, θ é o parâmetro pertencente ao espaço paramétrico Θ e $x \in R$.

   Nota-se que $J(f,g) = J(g,f)$ enquanto que $I(g,f) \neq I(f,g)$ a menos que $f = g$. Por esse motivo $J(g,f)$ é chamado de divergência simétrica de Kullback.

Neste estudo, focaremos na informação de Kullback-Leibler e na divergência simétrica de Kullback, que serão alvo de estudos dos critérios de informação. Segundo Emiliano (2013), a informação de Kullback-Leibler representa a informação perdida do modelo ajustado do modelo verdadeiro e muitos autores utilizam esse critério para medir a discrepância entre duas funções de probabilidade.

Conforme citado em Emiliano(2009), $I(g,f)$ tem as seguintes propriedades:

1. Quaisquer funções de densidade de probabilidade, g e f, $I(g,f) \geq 0$.
2. Se f e g são funções densidade de probabilidade e $I(g,f) = 0$, então $g(x) = f(x)$ para qualquer $x \in R$.
3. Dadas f e g, funções densidade de probabilidade com $f \to g$ então $I(g,f) \to 0$.

Seja $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ um conjunto de n observações amostradas aleatoriamente de uma distribuição (modelo) de probabilidades desconhecida $g(x)$, sendo realizações da variável aleatória $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ e $f(x)$ um modelo ajustado aos dados $\mathbf{x}$. Deseja-se então avaliar a qualidade do ajuste ao aproximar o modelo $g(x)$ pelo modelo $f(x)$. Para tal, ao comparar dois modelos candidatos $f_1(X|\theta_1)$ ou $f_2(X|\theta_2)$ para verificar qual modelo candidato melhor se aproxima do modelo $g(x)$, observa-se que o que diferencia os modelos são os segundos termos das equações (2.21) e (2.22), que são dadas por:

$$I(g;f_1) = E_g\left[ln\left(\frac{g(X)}{f_1(X)}\right)\right] = \int_{-\infty}^{+\infty} g(x)ln(g(x))dx - \int_{-\infty}^{+\infty} g(x)ln(f_1(x|\theta_1))dx \tag{2.21}$$

e

$$I(g;f_2) = E_g\left[ln\left(\frac{g(X)}{f_2(X)}\right)\right] = \int_{-\infty}^{+\infty} g(x)ln(g(x))dx - \int_{-\infty}^{+\infty} g(x)ln(f_2(x|\theta_2))dx \tag{2.22}$$

Vale notar que estes termos ainda dependem da distribuição desconhecida $g(x)$ e o termo que diferencia os dois modelos candidatos, $f_1(X|\theta_1)$ e $f_2(X|\theta_2)$, é o valor esperado da log verossimilhança e pode ser expressa como

$$E_g[ln(f(X))] = \int_{-\infty}^{+\infty} g(x)ln(f(x|\theta))dx. \tag{2.23}$$

Uma estimativa para o valor esperado se baseia na distribuição empírica dos dados, ou seja,

$$E_{\hat{g}}[ln(f(X))] = \frac{1}{n}\sum_{\alpha=1}^{n} ln(f(x_\alpha)). \tag{2.24}$$

De acordo com Konishi e Kitagawa (2008) tem-se que quando $n \to \infty$,

$$\frac{1}{n}\sum_{\alpha=1}^{n} ln(f(x_\alpha)) \to E_g[ln(f(X))] \tag{2.25}$$

Se multiplicarmos por n a equação (2.24) tem-se que

$$nE_{\hat{g}}[ln(f(X))] = \sum_{\alpha=1}^{n} ln(f(x_\alpha)) \tag{2.26}$$

que é o log verossimilhança da $f(x)$, ou seja, o log da função de verossimilhança é entendida como uma boa aproximação para a informação de Kullback-Leibler (Konishi e Kitagawa, 2008).

Uma boa estimativa do vetor de parâmetros, θ, de cada modelo candidato, pode ser obtida a partir do estimador de máxima verossimilhança, que apresentam ótimas propriedades assintóticas, tais como consistência, eficiência e suficiência. Logo é razoável aceitar

$$E_g[f(X|\hat{\theta})] = \int_{-\infty}^{+\infty} g(x) ln(f(x|\hat{\theta}))dx$$

Assim, tem-se que um estimador para $E_g[ln(f(X|\hat{\theta}))]$ é $n^{-1}l(\hat{\theta})$ e que $l(\hat{\theta})$ é um estimador de $nE_g[ln(f(X|\hat{\theta}))]$, com $l(\hat{\theta})$ sendo a função de log verossimilhança e $\hat{\theta}$ o estimador MV de θ.

2.7.1 Critérios de Informação

Konishi e Kitagawa (2008) afirmam que, embora o log da função de verossimilhança seja uma boa estimativa para a informação de Kullback-Leibler, ou seja, $l(\hat{\theta})$ é um bom estimador para $nE_g[log(f(X|\hat{\theta}))]$, o fato dos dados, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, serem utilizados duas vezes, ou seja, para a estimação dos parâmetros e para a estimação do valor esperado da log verossimilhança, introduz um viés na função de log verossimilhança, $l(\hat{\theta})$. Diante disso, os critérios de informação são construídos para avaliar e corrigir o viés da função suporte (Emiliano, 2009).

Seja $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ um conjunto de n observações amostradas aleatoriamente de uma distribuição (modelo) de probabilidades desconhecida $g(x)$, sendo realizações da variável aleatória $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ e seja

$$l(\theta) = \sum_{\alpha=1}^{n} ln(f(x_\alpha|\hat{\theta}(\mathbf{x}))) = ln(f(\mathbf{x}|\hat{\theta}(\mathbf{x}))$$

o log da função de verossimilhança, com $\hat{\theta}$ sendo o estimador MV de θ.

O viés da log verossimilhança como um estimador do valor esperado da log verossimilhança é

$$b = E_{g(\mathbf{x})}([ln(f(\mathbf{X}|\hat{\theta})) - nE_{g(Z)}[ln(f(Z|\hat{\theta}))]] \tag{2.27}$$

em que $E_{g(Z)}$ é o valor esperado da distribuição verdadeira $g(z)$.

Em geral, a forma de um critério de informação segue a característica

$$IC = -2ln(\textit{verossimilhança}) + 2\textit{viés} = \tag{2.28}$$

$$= -2\sum_{\alpha=1}^{n} ln(f(x_\alpha|\hat{\theta}(\mathbf{x}))) + 2\textit{viés}$$

.

O viés também depende da distribuição verdadeira $g(x)$ e portanto é necessário estimá-lo e os critérios de informação assumem diferentes penalizações.

1. Critério de informação de Akaike (AIC)

   O critério de informação de Akaike (AIC) ou também conhecido como mAIC (marginal AIC) no contexto do modelo linear misto (Vaida e Blanchard, 2005), baseia-se na informação de Kullback-Leibler e é muito conhecido e utilizado para selecionar modelos. Akaike(1974) mostrou que o viés introduzido pela estimação de máxima verossimilhança tende assintoticamente ao número de parâmetros a serem estimados no modelo. O Critério de Informação de Akaike é definido por:

$$AIC = -2\sum_{\alpha=1}^{n} ln(f(x_\alpha)|\hat{\theta}) + 2k \tag{2.29}$$

em que k representa o número de parâmetros do modelo.

Classifica-se como o melhor modelo candidato, no sentido que melhor se aproxima do modelo "verdadeiro"aquele modelo que apresente o menor valor de AIC.

2. Critério de informação Bayesiano (BIC)

O critério de informação Bayesiano também se baseia na informação de Kullback-Leibler e foi proposto por Schwarz (1978).

Conforme Konishi & Kitagawa (2008), assumindo r modelos candidatos, $M_1, M_2, \ldots, M_r$, em que cada modelo M_i é caracterizado por uma distribuição paramétrica, $f_i(x|\theta_i)$ e uma distribuição a priori, $\pi_i(\theta_i)$, em que o vetor parâmetro θ_i é k-dimensional. Seja $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ um conjunto de n observações, então a distribuição marginal de $\mathbf{x}$ para cada modelo M_i é dada por

$$p_i(\mathbf{x}) = \int f_i(\mathbf{x}|\theta_i)\pi_i(\theta_i)d\theta_i$$

Essa quantidade pode ser considerada como a verossimilhança para o modelo M_i.

Considerando $P(M_i)$ a probabilidade a priori do i-ésimo modelo, pelo teorema de Bayes, a probabilidade a posteriori é

$$P(M_i|\mathbf{x}) = \frac{p_i(\mathbf{x})P(M_i)}{\sum_{j=1}^{r} p_j(\mathbf{x})P(M_j)}$$

com $i = 1, 2, \ldots, r$.

Segundo Konishi e Kitagawa (2008), a probabilidade a posteriori indica a probabilidade dos dados serem gerados polo i-ésimo modelo quando os dados $\mathbf{x}$, são observados. Se um modelo está sendo selecionado dentre r modelos, adotar-se-á o modelo com maior probabilidade a posteriori.

Assim, BIC é definido por

$$BIC = -2lnf(\mathbf{x}|\hat{\theta}) + k\ \ ln(N) \tag{2.30}$$

com $\hat{\theta}$ sendo o estimador de MV de θ, k é o número de parâmetros do modelo e N é o número de observações.

Dessa forma, dentre os r modelos, classifica-se como o melhor modelo, aquele que apresente o menor valor de BIC.

3. Critério de informação Kullback (KIC)

O critério de informação de Kullback (KIC) baseia-se na divergência simétrica de Kullback (definida pela equação (2.20)) e foi definido por Cavanaugh (1999) como:

$$KIC = -2lnf(\mathbf{x}|\hat{\theta}) + 3k \tag{2.31}$$

em que k representa o número de parâmetros do modelo. Classifica-se como o melhor modelo, aquele que apresentar o menor valor de KIC.

2.8 *Software* R

O uso de softwares para análises estatísticas se tornou indispensável nos dias atuais. Com o avanço tecnológico, foi possível desenvolver softwares que realizassem processos rápidos e confiáveis. Um dos softwares utilizados neste trabalho é o software R.

O pacote *nlme* do R é uma ferramenta importante para a modelagem de dados de medidas repetidas, dados longitudinais e dados de curvas de crescimento. Neste trabalho será utilizado para os ajustes dos modelos mistos a função *lme* do pacote *nlme*. No anexo A está detalhada a forma de utilização da função *lme*.

Para apresentar os argumentos que compõem a função *lme* tem-se o seguinte exemplo:

```
fm1 <- lme(distance ~ Sex*age, weights = varIdent(form = ~ 1 | Sex),
correlation = corAR1(),data = Orthodont)
```

Utilizando o conjunto de dados Orthodont disponível no pacote *nlme* (especificado pelo argumento `data`), temos o modelo $fm1$ ajusta um modelo que considera diferença no intercepto e no coeficiente angular para o fator sexo (`distance`: variável resposta, `Sex`, `age`: covariáveis); considera também diferentes variâncias para cada sexo (definido pelo argumento `weights`) e uma estrutura de correlação AR1 (definido pelo argumento `correlation`). Por defaut, esse modelo é ajustado pelo MVR, que equivale a utilizar o argumento `method = REML`. Para utilizar o método MV basta inserir `method = ML`.

Alguns detalhes do uso do *software* R encontra-se em Anexos - Anexo A e algumas opções para o argumento `weights` e `correlation` estão disponíveis na tabela anexo2 e na tabela anexo3, respectivamente.

2.9 *Software* SAS

O SAS é um software que realiza análises estatísticas, e para análise de medidas repetidas e dados longitudinais disponibiliza o PROC MIXED.

A estrutura básica dos comandos usados no PROC MIXED é a seguinte:

```
PROC MIXED<opções>;
CLASS variáveis;
MODEL variável dependente=<efeitos fixos></opções>;
RANDOM efeitos aleatórios</opções>;
REPEATED efeitos repetidos</opções>;
```

O comando PROC MIXED inicia o procedimento de análise e algumas das suas principais opções são o DATA (especifica o arquivo com os dados de entrada) e `METHOD` (especifica o método de estimação). Nesta linha de comando podem ser inseridas outras opções para a saída de resultados e métodos de otimização dentre outros. No comando `CLASS` são declaradas as variáveis classificatórias ou qualitativas; no comando `MODEL` especifica-se o nome da variável resposta e a parte fixa do modelo misto, no comando `RANDOM` especifica-se a parte aleatória do modelo e escolhe-se a estrutura da matriz $\mathbf{G}$ e no `REPEATED`, a estrutura da matriz $\mathbf{R}_i$.

O comando `RANDOM` tem algumas opções importantes tais como `GROUP`=, `SUBJECT` = que especifica o nome fator relativo às unidades experimentais (sujeitos) que associa a cada elemento do fator especificado a mesma estrutura de covariâncias escolhida na opção `TYPE`.

O comando `REPEATED` tem algumas opções importantes tais como `GROUP`= que define a heterogeneidade dos parâmetros de covariâncias da matriz $\mathbf{R}$, `SUBJECT` = que especifica o sujeito onde são realizadas as medidas repetidas, `TYPE`= especifica a estrutura de covariância para a matriz $\mathbf{R}$. Algumas das estruturas para a matriz R podem sem encontradas na Tabela 2.1, em que t é o número de ocasiões.

Tabela 2.1. Algumas estruturas de covariância disponíveis no PROC MIXED do SAS

Estrutura	Descrição	Parâmetros	(i,j)-ésimo elemento
AR(1)	Autorregressiva (1)	2	$\sigma^2\rho^{\|i-j\|}$
ARH(1)	AR(1) heterogêneo	$t+1$	$\sigma_i\sigma_j\rho^{\|i-j\|}$
CS	Simetria Composta	2	$\sigma_1+\sigma^2 1(i=j)$
UN	Não Estruturada	$t(t+1)/2$	σ_{ij}
VC	Componente de Variância	q	$\sigma_k^2 1(i=j)$

Fonte: SAS/STAT (2008)

2.10 Avaliação e desempenho dos critérios de informação

Baseado na Tabela 2.2 admite-se que a decisão ideal ao se realizar um teste de hipótese é de não rejeitar a hipótese H_0 se de fato ela for verdadeira e de rejeitá-la quando ela de fato for falsa.

Tabela 2.2. Erros tipo I e tipo II envolvidos em um teste de hipóteses

	H_0 é verdadeira	H_0 é falso
Decisão de rejeitar H_0	Erro tipo I	Decisão correta
Decisão de não rejeitar H_0	Decisão correta	Erro tipo II

Uma maneira comum de encontrar um teste razoável consiste em fixar a probabilidade máxima do erro tipo I e tentar encontrar um teste com menor probabilidade do erro tipo II.

No contexto do estudo, considere a situação em que existem dois modelos $f(x)$ e $g(x)$ como geradores de n dados, $X_1, X_2, \ldots, X_n$. Suponha que os dados foram gerados verdadeiramente pelo modelo $g(x)$. Na escolha do modelo verdadeiro pode-se ter quatro situações, descritas a seguir:

Tabela 2.3. Classificação dos modelos

		Classe Prevista	
		sim	não
Classe Real	não	Falso Positivo (FP)	Verdadeiro Negativo (TN)
	sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)

1. Classificar $X_1, X_2, \ldots, X_n$ como originados do modelo $g(x)$ quando de fato estes originam-se do modelo $g(x)$. Classifica-se o modelo positivamente e tem-se um verdadeiro positivo (TP - *True Positive*).

2. Classificar $X_1, X_2, \ldots, X_n$ como não originados do modelo $g(x)$ quando estes se originam do modelo $g(x)$. Classifica-se incorretamente o modelo e tem-se um falso negativo (FN - *False Negative*).

3. Classificar $X_1, X_2, \ldots, X_n$ como originados do modelo $g(x)$ quando na verdade, se originam do modelo $f(x)$. Classifica-se o modelo incorretamente e tem-se um falso positivo (FP - *False Positive*).

4. Classificar $X_1, X_2, \ldots, X_n$ como não originados do modelo $g(x)$ quando de fato eles se originam do modelo $f(x)$. Classifica-se o modelo corretamente e tem-se um verdadeiro negativo (TN - *True Negative*).

Conforme descrito por Emiliano(2013) podem ser definidas as seguintes taxas:

1. Taxa verdadeiro positivo $TP(\%) = \dfrac{TP}{TP+FN}$ (fórmula), que fornece um indicativo da confiança, $(1-\alpha)$, do teste.

2. Taxa verdadeiro negativo $TN(\%) = \dfrac{TN}{TN + FP}$ (fórmula), que fornece um indicativo do poder do teste $(1 - \beta)$.

3. Taxa falso positivo $FP(\%) = \dfrac{FP}{FP + TN}$ indica(fórmula), que fornece um indicativo da probabilidade de cometer um erro do tipo II do teste (β).

4. Taxa falso negativo $FN(\%) = \dfrac{FN}{FN + TP}$ (fórmula), que fornece um indicativo da probabilidade de cometer um erro do tipo I do teste (α).

3 MATERIAL E MÉTODOS

Nesta seção será feita uma descrição do material utilizado no estudo de simulação e uma descrição de como este estudo foi realizado, utilizando os *softwares* R e SAS, além da definição dos cenários considerados no processo. No estudo de simulação será descrito o processo utilizando o *software* R, bem como os cenários a serem considerados na simulação.

3.1 Material

No presente estudo será utilizado um conjunto de dados de desempenho de frangos de corte, utilizado no trabalho de Lima (1988), envolvendo 32 frangos de corte da linhagem comercial Hubbard, sendo 13 fêmeas e 19 machos. As aves foram identificadas por uma anilha de alumínio e separadas por sexo ao primeiro dia de vida e alojadas em dois boxes separados. Foram alimentadas com a mesma ração comercial e cada ave foi pesada semanalmente, durante um período de 7 semanas, com pesagens feitas no mesmo dia da semana e no mesmo horário. Informações básicas sobre os dados são apresentadas na Tabela 3.1.

Um estudo desse conjunto de dados foi realizado no trabalho de Barbosa (2009) cujo objetivo foi encontrar um bom modelo que descrevesse os dados utilizando a abordagem por modelo linear misto. Para encontrar um bom modelo, foram utilizadas ferramentas gráficas e analíticas.

Tabela 3.1. Médias e erros-padrão (e.p.) dos pesos em gramas, de frangos de corte *Hubbard*, por sexo e semana de idade

	Fêmea		Macho	
Semana	Média	e.p.	Média	e.p.
1	133,30	36,42	129,53	65,00
2	321,38	81,71	324,26	120,56
3	559,92	200,02	621,79	179,23
4	807,54	255,97	895,42	286,15
5	1089,85	193,47	1241,58	514,22
6	1473,00	271,49	1702,47	582,65
7	1770,00	298,41	2067,37	714,59

Fonte: Barbosa (2009)

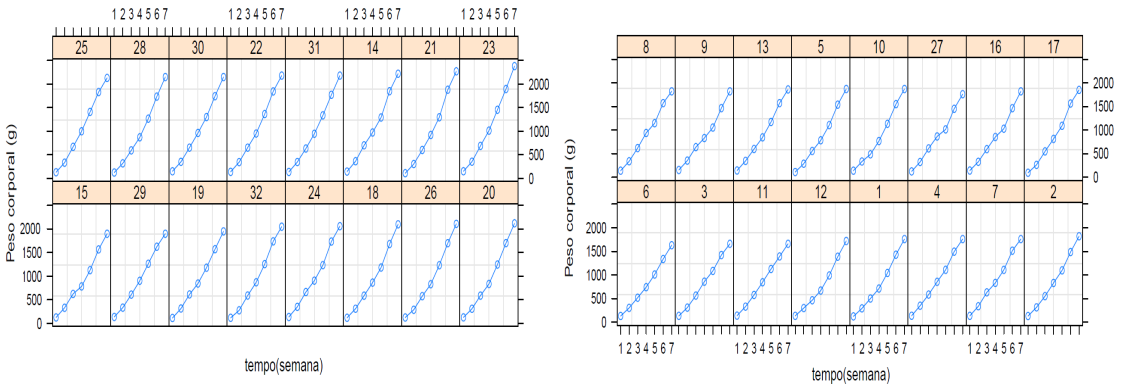

Figura 3.1. Perfis individuais dos pesos dos frangos ao longo do período experimental (1 ao 13: Fêmeas e 14 ao 32: Machos)

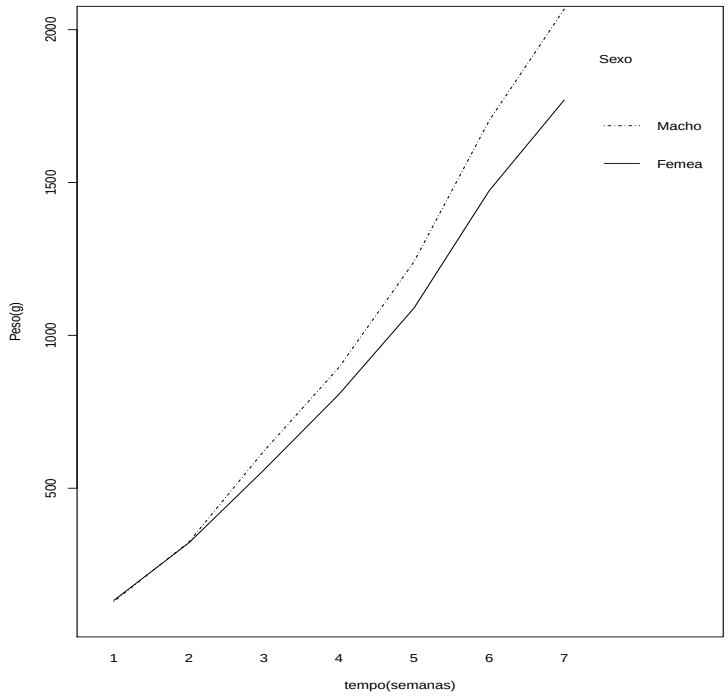

Figura 3.2. Perfís médios dos pesos dos frangos, separados por sexo.

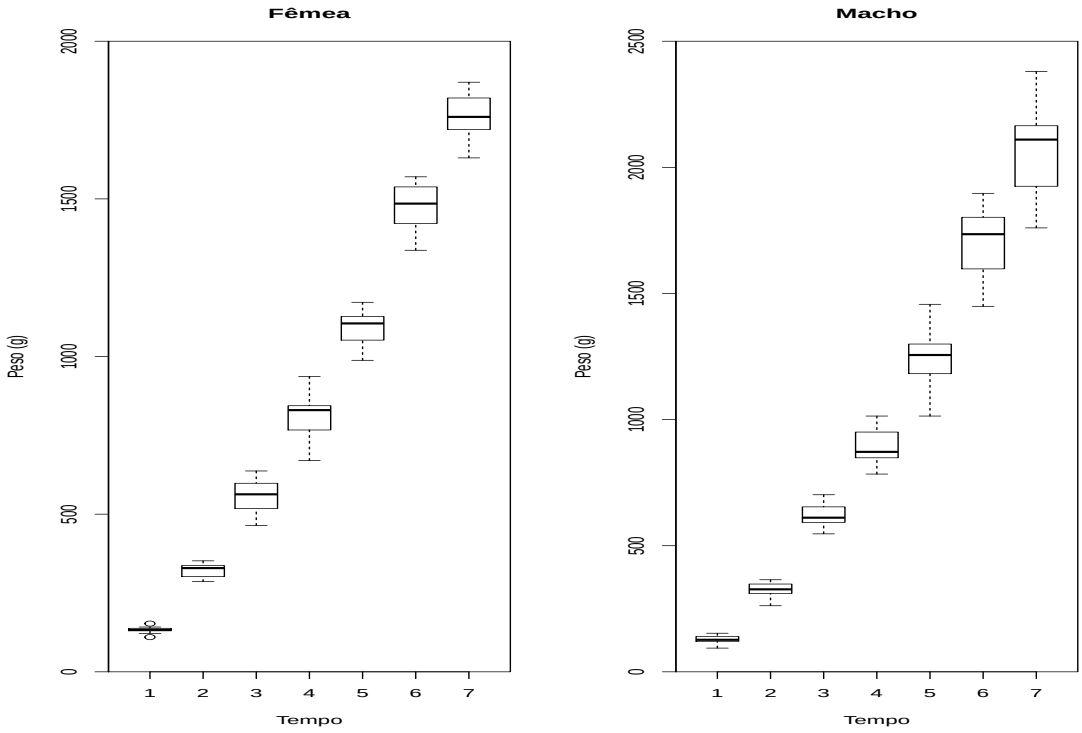

Figura 3.3. Gráfico de caixas dos pesos corporais dos frangos avaliados em 7 semanas, separados por sexo

Na matriz 3.1 os elementos da diagonal principal correspondem às estimativas das variâncias dos pesos nas 7 semanas, os elementos acima da diagonal principal são as covariâncias entre os pesos nas

diversas ocasiões e os elementos abaixo da diagonal, as correlações entre essas medidas.

$$\begin{bmatrix} \mathbf{435,62} & \mathbf{729,23} & \mathbf{943,45} & \mathbf{1248,72} & \mathbf{1365,23} & \mathbf{1147,27} & \mathbf{1499,69} \\ 0,85 & \mathbf{1661,23} & \mathbf{2353,61} & \mathbf{2897,68} & \mathbf{2827,15} & \mathbf{2746,97} & \mathbf{299,43} \\ 0,60 & 0,77 & \mathbf{5613,84} & \mathbf{6531,27} & \mathbf{6030,51} & \mathbf{6512,32} & \mathbf{6123,29} \\ 0,56 & 0,66 & 0,81 & \mathbf{11505,15} & \mathbf{12067,12} & \mathbf{12586,77} & \mathbf{2045,92} \\ 0,42 & 0,45 & 0,52 & 0,73 & \mathbf{23754,86} & \mathbf{24996,73} & \mathbf{27660,75} \\ 0,30 & 0,37 & 0,48 & 0,65 & 0,90 & \mathbf{32471,89} & \mathbf{36685,30} \\ 0,33 & 0,34 & 0,38 & 0,52 & 0,83 & 0,94 & \mathbf{47164,03} \end{bmatrix} \tag{3.1}$$

Na Tabela 3.1 e nas Figuras 3.1, 3.2 e 3.3 pode-se perceber o aumento dos pesos dos frangos ao longo do tempo, um grande aumento na variância dos pesos ao longo das semanas e uma diminuição no valor da correlação à medida que se aumenta o intervalo entre as avaliações. Em média, os machos tendem a ser mais pesados que as fêmeas a partir da terceira semana de vida.

A abordagem utilizada por Barbosa (2009) para encontrar um modelo linear misto que represente bem o crescimento em peso dos frangos foi: identificação dos efeitos aleatórios, escolha dos efeitos fixos e escolha da melhor estrutura de covariância entre e intra indivíduos (matrizes $\mathbf{G}$ e $\mathbf{R}_i$, respectivamente).

As sugestões para inclusão de efeito aleatório foram baseadas no gráfico de perfis individuais (Figura 3.1), em que os perfis do 1 ao 13 se referem às fêmeas e do 14 ao 32 aos machos. O grau do polinômio que explica bem o crescimento dos frangos baseou-se no comportamento dos perfis médios de peso (Figura 3.2) e as possíveis estruturas de covariâncias foram sugeridas pela matriz de estimativas de variâncias, covariâncias e correlações apresentada na expressão (3.1).

Inicialmente, Barbosa (2009) considerou o modelo maximal que incluiu o efeito fixo de *tempo*, *tempo*^{2}, *sexo*, interação de *sexo* e *tempo* e interação de *sexo* e *tempo*^{2} e efeito aleatório no intercepto, no termo linear e no termo quadrático, que pode ser escrito como:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 M_i + \beta_4 M_i t_{ij} + \beta_5 M_i t_{ij}^2 + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij} \tag{3.2}$$

em que $i = 1, \ldots, 32$ e $j = 1, \ldots, 7$, $M_i = 0$ se for fêmea e $M_i = 1$ se for macho. O modelo (3.2) pode ser escrito como:

$$y_{ij} = \begin{cases} \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij} & \text{, se for Fêmea} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_4) t_{ij} + (\beta_2 + \beta_5) t_{ij}^2 + b_{0i} + b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij} & \text{, se for Macho} \end{cases} \tag{3.3}$$

Na notação matricial, $\mathbf{b}_i = (b_{0i}, b_{1i}, b_{2i})'$ com $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$, considerou a matriz $\mathbf{G}$ como não estruturada e a matriz $\mathbf{R}_i = \sigma^2 \mathbf{I}$.

A seguir, a autora comparou o modelo (3.2) com outros modelos mais simples (modelos aninhados) com o intuito de escolher o melhor conjunto de efeitos aleatórios utilizando o teste de razão de verossimilhanças e os critérios AIC e BIC. Após a escolha dos efeitos aleatórios, buscou a melhor estrutura de efeitos fixos do modelo utilizando os testes t e Wald. Definidas as partes fixa e aleatória do modelo, prosseguiu com a escolha da melhor estrutura para a matriz de covariâncias intra-indivíduos ($\mathbf{R}_i$), utilizando o teste da razão de verossimilhanças e os critérios AIC e BIC.

Os modelos considerados por Barbosa (2009) são descritos a seguir:

Tabela 3.2. Relação de modelos lineares mistos utilizados na análise do peso corporal dos frangos de corte (Barbosa, 2009)

Modelo	Método Estimação	Efeito Fixo	Efeito Aleatório	Estrutura $\mathbf{R}_i$
F1	MVR	β_0, β_1, β_2 e β_4	*intercepto*, *tempo*, $tempo^2$	VC
F2	MVR	β_0, β_1, β_2 e β_4	*tempo* e $tempo^2$	VC
F2.1	MV	β_0, β_1, β_2 e β_4	*tempo* e $tempo^2$	VC
F3	MVR	β_0, β_1, β_2 e β_4	*intercepto* e *tempo*	VC
F4	MV	β_0, β_1, β_2, β_3 e β_5	*tempo* e $tempo^2$	VC
F5	MV	β_0, β_1, β_2 e β_5	*tempo* e $tempo^2$	VC
F6	MV	β_0, β_1, β_2 e β_5	(-)	UN
F6.1	MVR	β_0, β_1, β_2 e β_5	(-)	UN
F7	MV	β_0, β_1, β_2 e β_5	*tempo* e $tempo^2$	ARH(1)
F8	MV	β_0, β_1, β_2 e β_5	*tempo* e $tempo^2$	AR(1)
F9	MV	β_0, β_1, β_2 e β_5	*tempo* e $tempo^2$	CS

Fonte: Adaptado de Barbosa (2009)
Nota: (-) sem efeito aleatório
MV - Máxima Verossimilhança
MVR - Máxima Verossimilhança Restrita
VC - Componente de Variância
UN - Não Estruturada
ARH(1) - Auto Regressiva com Heterogeneidade de Variância
AR(1) - Auto Regressiva de ordem 1
CS - Simetria Composta

De acordo com o critério de informação BIC utilizado na comparação de modelos, Barbosa (2009) concluiu que o modelo F7 foi o mais adequado para descrever os dados dos pesos dos frangos. Este modelo foi utilizado como o modelo verdadeiro que gerou os dados e que pode ser escrito como:

$$y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_5 M_i t_{ij}^2}_{\text{parte fixa}} + \underbrace{b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij}}_{\text{parte aleatória}} \tag{3.4}$$

com $i = 1, 2, \ldots, 32$ e $j = 1, 2, \ldots, 7$ e as suposições $\mathbf{b}_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{G})$, $\mathbf{G} = cov(\mathbf{b}_i) = \begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1,b_2} \\ \sigma_{b_1,b_2} & \sigma^2_{b_2} \end{bmatrix}$ e $\mathbf{R}_i = cov(\varepsilon_{ij}) = \text{ARH(1)}$.

3.1.1 Estudo de Simulação

O modelo (3.4) foi utilizado para simular os pesos de frangos. A partir dos dados gerados será analisado o comportamento dos critérios de informação em situações que serão descritas a seguir.

3.1.1.1 Descrição da simulação

1. R

Para a simulação dos pesos dos frangos, considerando as características do modelo (3.4), foi utilizado o *software* R seguindo os passos:

Passo 1: Definições do número de repetições (j), número de indivíduos (m), valores dos parâmetros β;

Passo 2: Construção da matriz de delineamento $\mathbf{X}$;

Passo 3: Construção da matriz $\mathbf{Z}$;

Passo 4: Definição da matriz $\mathbf{G}$ (No nosso trabalho a estrutura da matriz $\mathbf{G}$ é a Não Estruturada (UN));

Passo 5: Gerar valores aleatórios para o vetor **b**;

Passo 6: Construção da estrutura de covariancia $\mathbf{R}_i$;

Passo 7: Construção do vetor **y** (vetor dos pesos das aves);

Depois de obter o vetor **y** prossegue-se ajustando os modelos candidatos:

Passo 8: Construção do vetor de indivíduos, fator sexo e do objeto `groupedData`;

Passo 9: Ajuste dos modelos candidatos;

Passo 10: Armazena o menor valor de AIC, BIC e KIC;

Passo 11: Repete-se as etapas anteriores 1000 vezes.

O código que detalha os processos descritos para a simulação dos pesos dos frangos se encontram no apêndice.

2. SAS

O *software* adotado para este estudo foi o R, porém Wincklin (2013) apresenta os procedimentos para a simulação do vetor **y** com as características de um modelo misto. O autor considerou as seguintes etapas para o processo de simulação:

Passo 1: Construir as matrizes de delineamento **X** e **Z** considerando o delineamento dos dados. Utilizar o procedimento GLIMMIX.

Passo 2: Construir a matriz **G**.

Passo 3 : Construir a matriz **R**.

Passo 4 : Construir a variável resposta (peso) somando os termos

a) $\mathbf{X}\beta$

b) $\mathbf{Z}b$, em que $b \sim MVN(\mathbf{0}, \mathbf{G})$

c) ε, em que $\varepsilon \sim MVN(\mathbf{0}, \mathbf{R})$.

O código detalhado do procedimento de simulação do modelo misto utilizando o *software* SAS encontra-se em Anexos - Anexo B.

Cabe ressaltar que neste trabalho, as análises da performance dos critérios de informação, foram baseadas nas simulações dos pesos dos frangos feitos no *software* R.

3.1.1.2 Seleção dos efeitos fixos

Considerou-se que o modelo (3.4) representa o modelo verdadeiro, exceto pela estrutura da matriz $\mathbf{R}_i$, ou seja,

$$y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_5 M_i t_{ij}^2}_{\text{parte fixa}} + \underbrace{b_{1i} t_{ij} + b_{2i} t_{ij}^2 + \varepsilon_{ij}}_{\text{parte aleatória}}$$

com as suposições de $\mathbf{b}_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{G})$, $\mathbf{G} = cov(\mathbf{b}_i) = \begin{bmatrix} \sigma_{b_1}^2 & \sigma_{b_1,b_2} \\ \sigma_{b_1,b_2} & \sigma_{b_2}^2 \end{bmatrix}$ e $\mathbf{R}_i = cov(\varepsilon_{ij}) = \sigma^2\mathbf{I}$.

Para estudar o desempenho dos critérios de informação (AIC, BIC e KIC) na seleção da estrutura de efeitos fixos, foram utilizadas as estimativas dos parâmetros de efeito fixo e os parâmetros de covariâncias das matrizes **G** do modelo escolhido por Barbosa (2009) (tabela 3.3) e $\mathbf{R}_i = \sigma^2\mathbf{I}$, no processo de simulação dos pesos de frangos, considerando $m = 32$ frangos e $n_i = n = 7$ tempos.

Tabela 3.3. Valores dos parâmetros utilizados na simulação dos pesos de frangos

Parâmetro	Valor
β_0	-48.0491
β_1	152.9825
β_2	15.6035
β_5	6.1909
$\sigma^2_{b_1}$	311.0990
σ_{b_1,b_2}	-44.1101
$\sigma^2_{b_2}$	13.9110
σ^2	1441.8068

Para o conjunto de dados simulados, foram ajustados pelo método MV os seis modelos candidatos listados a seguir, que têm os mesmos efeitos aleatórios, estruturas de covariâncias conhecidas e que diferem entre si somente na parte fixa.

Tabela 3.4. Modelo verdadeiro (MF6) e modelos candidatos usados no ajuste do peso de frangos - parte fixa

Modelo	Parte fixa
MF1	$\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 M_i + \beta_4 M_i t_{ij} + \beta_5 M_i t_{ij}^2$
MF2	$\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 M_i + \beta_5 M_i t_{ij}^2$
MF3	$\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 M_i + \beta_4 M_i t_{ij}$
MF4	$\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2$
MF5	$\beta_0 + \beta_2 t_{ij}^2 + \beta_5 M_i t_{ij}^2$
MF6*	$\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_5 M_i t_{ij}^2$

*MF6: modelo verdadeiro

Dentre os seis modelos escolheu-se o modelo com menor valor de AIC, BIC e KIC. Este processo foi repetido 1000 vezes e quantificou-se a taxa de acerto de cada um dos critérios de informação. Vale notar que os a taxa de acerto de cada critério de informação contabiliza a proporção de vezes que o critério (AIC, BIC e KIC) indica o modelo MF6 como sendo o modelo correto.

Alguns desses modelos candidatos foram ajustados no trabalho de Barbosa (2009), tais como o modelo MF1 e MF2, mas o modelo MF6 foi considerado o melhor pela autora e por isso esse modelo foi utilizado para a simulação do conjunto de dados. Outros modelos candidatos foram considerados neste trabalho como possíveis modelos a serem escolhidos pelos critérios de informação.

A fim de avaliar o comportamento dos critérios de informação sob diferentes tamanhos de amostra, neste trabalho foram considerados também simulações envolvendo um número menor ($m = 10$, 4 fêmeas e 6 machos) e outro maior de aves ($m = 50$, 20 fêmeas e 30 machos).

3.1.1.3 Seleção dos efeitos aleatórios

Para estudar o desempenho dos critérios de informação (AIC, BIC e KIC) na seleção da estrutura de efeitos aleatórios, foram utilizadas as estimativas dos parâmetros de efeito fixo e os parâmetros de covariâncias das matrizes $\mathbf{G}$ do modelo escolhido por Barbosa (2009) (tabela 3.3) e $\mathbf{R}_i = \sigma^2\mathbf{I}$, no processo de simulação dos pesos de frangos, considerando $m = 32$ frangos e $n_i = n = 7$ tempos.

Tabela 3.5. Valores dos parâmetros utilizados na simulação dos pesos de frangos

Parâmetro	Valor
β_0	-48.0491
β_1	152.9825
β_2	15.6035
β_5	6.1909
$\sigma^2_{b_1}$	325.50
σ_{b_1,b_2}	-46.2661
$\sigma^2_{b_2}$	14.7671
σ^2	1450.9200

A partir dos dados simulados, foram ajustados 7 modelos candidatos (apresentados na Tabela 3.6) pelo método da máxima verossimilhança restrita (MVR), admitindo-se a parte fixa do modelo e a estrutura da matriz $\mathbf{R}_i$ são conhecidas.

Tabela 3.6. Modelo verdadeiro (MA7) e modelos candidatos usados no ajuste do peso de frangos - parte aleatória

Modelo	Parte aleatória
MA1	$b_{0i} + b_{1i}t_{ij} + b_{2i}t^2_{ij}$
MA2	b_{0i}
MA3	$b_{1i}t_{ij}$
MA4	$b_{2i}t^2_{ij}$
MA5	$b_{0i} + b_{1i}t_{ij}$
MA6	$b_{0i} + b_{2i}t^2_{ij}$
MA7*	$b_{1i}t_{ij} + b_{2i}t^2_{ij}$

*MA7: modelo verdadeiro

Dentre os sete modelos candidatos, foi considerado como o melhor modelo aquele que apresentou o menor valor para AIC, BIC e KIC. Este processo foi repetido 1000 vezes e foi analisado o desempenho de cada critério de informação.

A fim de avaliar o comportamento dos critérios de informação sob diferentes tamanhos de amostra, neste trabalho foram considerados também simulações envolvendo um número menor ($m = 10$, 4 fêmeas e 6 machos) e outro maior de aves ($m = 50$, 20 fêmeas e 30 machos).

No estudo de seleção da estrutura simetria composta (CS) foi considerado a situação em que a matriz de covariância $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$ assume essa estrutura. Para tal, foi considerado efeito aleatório no intercepto e a estrutura de $\mathbf{R}_i = \sigma^2\mathbf{I}$. Com o intuito de selecionar os efeitos aleatórios, considera-se que os efeitos fixos e a matriz $\mathbf{R}_i = \sigma^2\mathbf{I}$ são conhecidos.

Na simulação dos pesos dos frangos considerando essa estrutura, foram utilizados os seguintes valores de parâmetros, que são as estimativas dos parâmetros obtidas no trabalho de Barbosa (2009):

Tabela 3.7. Valores dos parâmetros utilizados na simulação dos pesos de frangos

Parâmetro	Valor
β_0	-48.0491
β_1	152.9825
β_2	15.6035
β_5	6.1909
$\sigma^2_{b_0}$	3135.181
σ^2	4137.172

A partir do conjunto de dados gerado, foram ajustados seis modelos candidatos, considerando que a parte fixa do modelo e a estrutura da matriz $\mathbf{R}_i$ são conhecidas.

Tabela 3.8. Modelo verdadeiro (MS1) e modelos candidatos usados no ajuste do peso de frangos - parte aleatória

Modelo	Parte aleatória
MS1*	b_{0i}
MS2	$b_{1i}t_{ij}$
MS3	$b_{2i}t_{ij}^2$
MS4	$b_{0i} + b_{1i}t_{ij}$
MS5	$b_{0i} + b_{2i}t_{ij}^2$
MS6	$b_{1i}t_{ij} + b_{2i}t_{ij}^2$

*MS1: modelo verdadeiro

Dentre os seis modelos candidatos, foi considerado como o melhor modelo aquele que apresentou o menor valor para AIC, BIC e KIC. Este processo foi repetido 1000 vezes e foi analisado o desempenho de cada critério de informação.

A fim de avaliar o comportamento dos critérios de informação sob diferentes tamanhos de amostra, neste trabalho foram considerados também simulações envolvendo um número maior de aves ($m = 50$, 20 fêmeas e 30 machos).

3.1.1.4 Seleção da estrutura de covariância $\mathbf{R}_i$

Neste estudo, a parte fixa do modelo e a estrutura da matriz $\mathbf{G}$ foram consideradas como conhecidas, como apresentado no modelo (3.4). O objetivo desse estudo é avaliar o comportamento dos critérios de informação na seleção da estrutura de covariância do modelo misto.

Para a simulação foram consideradas algumas estruturas para a matriz $\mathbf{R}_i$, e seus respectivos valores dos parâmetros são apresentados na Tabela 3.9. Os valores dos parâmetros utilizados na simulação dos dados são as estimativas dos parâmetros encontrados no trabalho de Barbosa (2009).

Tabela 3.9. Valores dos parâmetros utilizados na simulação dos pesos de frangos

Estrutura	Termo geral	Componentes de $\mathbf{R}_i$
Diagonal	$\sigma_i^2 = \sigma^2, \sigma_{ij} = 0$	$\sigma^2 = 1450.92$
AR(1)	$\sigma^2\rho^{\lvert i-j\rvert}$	$\sigma^2 = 1450.92$, $\rho = -0.0329$ ($\rho = 0.5$, $\rho = 0.8$)
ARH(1)	$\sigma_{ij} = \sigma_i^2\sigma_j^2\rho^{\lvert i-j\rvert}$	$\sigma_1^2 = 62.72$, $\sigma_2^2 = 776.51$, $\sigma_3^2 = 6847.28$, $\sigma_4^2 = 7331.15$, $\sigma_5^2 = 8528.2$, $\sigma_6^2 = 21495$, $\sigma_7^2 = 8311.17$, $\rho = 0.8640$

Foram ajustados quatro modelos, pelo método da máxima verossimilhança restrita (MVR), sendo que o modelo MD1, MD2 e MD3 da Tabela 3.10 representam respectivamente os modelos verdadeiros da estrutura diagonal, AR(1) e ARH(1).

Tabela 3.10. Modelo verdadeiro e modelos candidatos usados no ajuste do peso de frangos - estrutura de covariância $\mathbf{R}_i$

Modelo	Estrutura $\mathbf{R}_i$
MD1	Componente de Variância (VC)
MD2	AR(1)
MD3	ARH(1)
MD4	Simetria Composta (CS)

Dentre os quatro modelos escolhe-se o modelo com menor valor de AIC, BIC e KIC. Este processo foi repetido 1000 vezes e verificou-se qual a taxa de acerto de cada um dos critérios de informação.

A fim de avaliar o comportamento dos critérios de informação sob diferentes tamanhos de amostra, neste trabalho foram considerados simulações envolvendo um número maior de aves ($m = 50$, 20 fêmeas e 30 machos). Na avaliação dos critérios de informação na seleção da estrutura AR(1) foram consideradas outros valores do parâmetro de correlação, pois o valor utilizado na simulação (estimativa obtida no trabalho de Barbosa (2009)) é um valor próximo de zero (os efeitos aleatórios nos termos linear e quadrático já explicam as correlações existentes entre as medidas repetidas). Diante disso foram considerados $\rho = 0.5$ e $\rho = 0.8$, valores estes que caracterizam mais a estrutura AR(1).

4 RESULTADOS E DISCUSSÕES

Nesta seção serão apresentados e discutidos os resultados obtidos do estudo de simulação descrito na seção 3.

4.1 Efeitos fixos

No estudo de seleção de efeitos fixos, o objetivo foi de avaliar os critérios de informação AIC, BIC e KIC na seleção dos efeitos fixos do modelo, ou seja, foi considerado que os efeitos aleatórios e a estrutura de covariância da matriz $\mathbf{R}_i$ era conhecida.

Para avaliar qual critério de informação tem o melhor desempenho nessa situação, foi descrito na tabela 4.1 o número de vezes em que cada modelo foi selecionado pelos critérios de informação, ou seja, para cada amostra foram calculados os valores dos critérios de informação para cada modelo e aquele que teve o menor valor do critério de informação, o modelo foi contabilizado; e as taxas de verdadeiro positivo que quantifica o quanto os critérios de informação escolheram o modelo verdadeiro.

Tabela 4.1. Número de indicações de cada modelo (parte fixa) por critério de informação, considerando amostras de 10, 32 e 50 aves

Amostra	Critério de Informação	MF1	MF2	MF3	MF4	MF5	MF6*	TP(%)
m=10	AIC	56	157	116	25	0	646	64.6 %
	BIC	6	48	45	133	0	768	76.8 %
	KIC	22	93	79	64	0	742	74.2 %
m=32	AIC	60	165	25	0	0	750	75.0 %
	BIC	4	23	2	0	0	971	97.1 %
	KIC	21	96	18	0	0	865	86.5 %
m=50	AIC	72	153	9	0	0	766	76.6 %
	BIC	1	17	4	0	0	978	97.8 %
	KIC	25	84	10	0	0	881	88.1 %

*MF6: modelo verdadeiro

De acordo com os resultados obtidos pode-se perceber que os critérios de informação apresentaram melhor desempenho em situações em que o tamanho da amostra é maior. Quando foi considerado $m = 50$ frangos, todos os critérios de informação tiveram um desempenho melhor em relação a $m = 10$ e $m = 32$ frangos.

Observa-se também que em todas as situações, o critério de informação BIC teve um desempenho melhor do que os critérios de informação AIC e KIC.

4.2 Efeitos aleatórios

No estudo de seleção de efeitos aleatórios, o objetivo foi de avaliar o desempenho dos critérios de informação AIC, BIC e KIC na seleção dos efeitos aleatórios do modelo, considerando que os efeitos fixos e a estrutura de covariância da matriz $\mathbf{R}_i$ eram conhecidas.

Para avaliar qual critério de informação tem o melhor desempenho nessa situação, foram descritos na Tabela 4.2 o número de vezes em que cada modelo foi selecionado pelos critérios de informação, ou seja, para cada amostra foi calculado os valores dos critérios de informação de cada modelo e aquele que teve o menor valor do critério de informação, o modelo foi contabilizado; e as taxas de verdadeiro positivo que considera que o critério de informação que teve um melhor desempenho é aquele que teve a maior taxa de verdadeiro positivo, ou seja, aquele critério de informação que mais acertou o modelo verdadeiro.

Tabela 4.2. Número de indicações e taxas de verdadeiro positivo (TP) de cada modelo (efeitos aleatórios) por critério de informação, considerando amostras de 10, 32 e 50 aves

Amostra	Critério de Informação	MA1	MA2	MA3	MA4	MA5	MA6	MA7*	TP(%)
	AIC	-	0	54	142	87	202	515	51.5 %
m=10	BIC	-	0	118	293	57	133	399	39.9 %
	KIC	-	0	74	219	76	167	464	46.4 %
	AIC	368	0	0	4	10	79	539	53.9 %
m=32	BIC	58	0	0	35	19	113	775	77.5 %
	KIC	209	0	0	13	13	100	665	66.5 %
	AIC	406	0	0	0	0	27	567	56.7 %
m=50	BIC	41	0	0	4	0	57	898	89.8 %
	KIC	249	0	0	0	0	40	711	71.1 %

*MA7: modelo verdadeiro

Com base nas taxas de acerto dos critérios de informação, tem-se que os critérios de informação melhoram suas performances com o aumento do tamanho da amostra. Para $m = 50$ e $m = 32$ aves o critério BIC teve um desempenho melhor que os demais, porém quando se considera $m = 10$ aves, BIC teve o pior desempenho.

De um modo geral, percebe-se que os critérios de informação não apresentaram um bom desempenho na seleção dos efeitos aleatórios.

Quanto a seleção de efeitos aleatórios considerando que a matriz $\mathbf{V}_i$ assume estrutura simetria composta (CS), obteve-se os seguintes resultados:

Tabela 4.3. Número de indicações e taxas de verdadeiro positivo (TP) de cada modelo (efeitos aleatórios ($\mathbf{V}_i$ =CS)) por critério de informação, considerando amostras de 32 e 50 aves

Amostra	Critério de Informação	MS1*	MS2	MS3	MS4	MS5	MS6	TP(%)
	AIC	750	0	0	94	156	0	75.0 %
m=32	BIC	953	0	0	20	27	0	95.3 %
	KIC	846	0	0	61	93	0	84.6 %
	AIC	779	0	0	86	135	0	77.9 %
m=50	BIC	958	0	0	17	25	0	95.8 %
	KIC	862	0	0	52	86	0	86.2 %

*MS1: modelo verdadeiro

Na situação em que a matriz $\mathbf{V}_i$ tem estrutura de Simetria Composta (CS), os critérios de informação tiveram um desempenho similar quando o tamanho das amostras é de 32 e 50 aves. Dentre os três critérios de informação, BIC teve o melhor desempenho nessa situação.

4.3 Simulação e escolha da estrutura de covariâncias $\mathbf{R}_i$

No estudo de seleção da estrutura de covariâncias $\mathbf{R}_i$, o objetivo foi de avaliar o desempenho dos critérios de informação AIC, BIC e KIC na seleção de algumas estruturas de covariância $\mathbf{R}_i$ considerando que os efeitos fixos e a estrutura de covariância da matriz $\mathbf{R}_i$ eram conhecidas, ou seja, a mesma do modelo (3.4) .

Nesta seção serão apresentados os resultados obtidos na simulação dos pesos dos frangos considerando diferentes estruturas de covariância da matriz $\mathbf{R}_i$.

1. Estrutura Diagonal

Na Tabela 4.4 é apresentada o número de escolhas de cada modelo pelos critérios de informação e a taxa de verdadeiro positivo (TP), que quantifica o acerto do modelo verdadeiro, considerando amostras de 32 e 50 aves.

Tabela 4.4. Número de indicações e taxas de verdadeiro positivo (TP) dos critérios de informação na escolha da estrutura $\mathbf{R}_i$ = VC considerando amostras 32 e 50 aves

Amostra	Critério de Informação	MD1*	MD2	MD3	MD4	TP(%)
m = 32	AIC	735	130	46	89	73.5 %
	BIC	972	18	0	10	97.2%
	KIC	882	77	4	37	88.2%
m = 50	AIC	710	129	48	113	71.0%
	BIC	978	15	0	7	97.8%
	KIC	871	76	4	49	87.1%

*MD5: modelo verdadeiro

Neste cenário, os critérios de informação tiveram desempenhos similares quando o tamanho da amostra é de 32 e 50 aves. Dentre os três critérios de informação, BIC se desempenhou melhor.

2. Estrutura AR(1)

Na Tabela 4.5 é apresentado o número de vezes em que os critérios de informação escolheram cada modelo e a taxa de acerto de cada critério de informação considerando amostras de 32 e 50 aves com diferentes valores do parâmetro ρ.

Tabela 4.5. Número de indicações e taxas de verdadeiro positivo (TP) dos critérios de informação na escolha da estrutura $\mathbf{R}_i$ = AR(1) para alguns valores do parâmetro ρ e diferentes tamanhos de amostra

Amostra	ρ	Critério de Informação	MD1	MD2*	MD3	MD4	TP(%)
m = 32	$\rho = -0.0329$	AIC	711	156	55	78	15.6 %
		BIC	958	36	0	6	3.6 %
		KIC	864	97	5	34	9.9 %
	$\rho = 0.5$	AIC	8	876	92	24	87.6 %
		BIC	58	918	0	24	91.8 %
		KIC	15	947	11	27	94.7 %
	$\rho = 0.8$	AIC	0	889	100	11	88.9 %
		BIC	0	987	0	13	98.7 %
		KIC	0	971	16	13	97.1 %
m = 50	$\rho = 0.5$	AIC	0	926	67	7	92.6 %
		BIC	10	983	0	7	98.3 %
		KIC	1	986	6	7	98.6 %
	$\rho = 0.8$	AIC	0	906	92	2	90.6%
		BIC	0	997	0	3	99.7 %
		KIC	0	990	7	3	99.0 %

*MD2: modelo verdadeiro

Quando o valor do parâmetro de correlação ρ considerado para a simulação foi de -0.0329, todos os critérios de informação (AIC, BIC e KIC) não se desempenharam bem. Isso se deve ao fato do valor de correlação ser próxima de zero, ou seja, quase não há presença de correlação entre as medidas repetidas, explicando assim, a preferência dos critérios de informação na escolha do modelo MA2, que se trata do modelo cuja estrutura de covariância é a de componentes de variância (VC).

Nas situações em que $\rho = 0.5$ e $\rho = 0.8$, os critérios de informação BIC e KIC tiveram desempenhos similares em ambas as amostras (32 e 50 aves), e tiveram um desempenho melhor que o critério AIC.

3. Estrutura ARH(1)

Na Tabela 4.6 é apresentado o número de vezes em que os critérios de informação escolheram cada modelo e a taxa de acerto de cada critério de informação considerando amostras de 32 e 50 aves.

Tabela 4.6. Número de indicações e taxas de verdadeiro positivo (TP) dos critérios de informação na escolha da estrutura $\mathbf{R}_i$ = ARH(1) considerando amostras de 32 e 50 aves

Amostra	Critério de Informação	MD1	MD2	MD3*	MD4	TP(%)
m = 32	AIC	0	0	1000	0	100.0 %
	BIC	0	0	1000	0	100.0 %
	KIC	0	0	1000	0	100.0 %
m = 50	AIC	0	0	1000	0	100.0 %
	BIC	0	0	1000	0	100.0 %
	KIC	0	0	1000	0	100.0 %

*MD3: modelo verdadeiro

De acordo com os resultados apresentados na Tabela 4.6 concluiu-se que o tamanho da amostra não tem impacto na performance dos critérios de informação considerando o cenário obtido pelo estudo de Barbosa (2009). Nesta situação, todos os critérios de informação tiveram um ótimo desempenho, tendo 100% de acerto na escolha do modelo verdadeiro.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, estudou-se o desempenho dos critérios de informação, AIC, BIC e KIC, em diferentes cenários simulados, comparando as suas taxas de verdadeiro positivo (taxa TP) em conjuntos de dados simulados utilizando o software R.

Nos estudos de seleção de efeitos fixos pode-se concluir que todos os critérios apresentaram melhores desempenhos quando o tamanho da amostra foi maior. O critério BIC teve melhor desempenho, seguido pelos critérios KIC e AIC (que apresentou o pior desempenho). De um modo geral, todos os critérios desempenham-se bem na seleção de efeitos fixos do modelo, concordando com os resultados obtidos nos trabalhos de Abraham (2008) e Gurka (2006).

De um modo geral, nos estudos de seleção de efeitos aleatórios os critérios de informação apresentaram desempenhos piores do que nos estudos de seleção dos efeitos fixos. Neste contexto, o critério BIC apresentou desempenho melhor que os demais, com exceção do cenário envolvendo o menor tamanho de amostra (10 aves), quando foi o pior deles. Quando se considerou o efeito aleatório somente no intercepto, os três critérios tiveram praticamente o mesmo desempenho, em amostras de tamanho 32 e 50.

Nos estudos de escolha da estrutura de covariâncias $\mathbf{R}_i$, pode-se concluir que nos cenários envolvendo amostras de 32 ou 50 aves, todos os critérios apresentaram desempenhos muito bons, sendo o critério BIC, o que apresentou maior TP, ou seja, indicou a estrutura verdadeira com maior frequência. No caso da estrutura AR(1), com parâmetro de correlação serial próximo a zero, todos os critérios apresentaram péssimos desempenhos, indicando menos de 20% das vezes a estrutura verdadeira, e quando o parâmetro de correlação serial foi considerado igual a 0.5 e 0.8, os critérios de informação BIC e KIC tiveram desempenhos similares. Já no caso da estrutura ARH(1), todos os critérios apresentaram ótimos desempenhos para os dois tamanhos de amostra. Neste estudo o critério BIC teve um desempenho melhor em relação ao AIC e KIC, não indicando o mesmo resultado obtido por Abraham (2008), já que o critério AIC foi considerado o melhor na seleção de estruturas mais complexas e o critério BIC em estruturas mais simples.

Baseado nos resultados de simulação de diversos cenários, comuns em experimentos de crescimento de animais, pode-se concluir que o critério BIC apresentou melhor desempenho que os critérios AIC e KIC, sendo considerado ideal para indicar a escolha do melhor modelo linear misto em situações semelhantes às que foram apresentados neste trabalho.

REFERÊNCIAS

ABRAHAM, A. A., 2008 *Model Selection Methods in the Linear Mixed Model for Longitudinal Data*. Master's thesis, University of North Carolina at Chapel Hill.

AKAIKE, H., 1974 A new look at the statistical model identification. IEEE Transactions on automatic control **19**: 716–723.

ALCARDE, R., 2012 *Modelos lineares mistos em dados longitudinais com o uso do pacote ASReml - R*. Ph.D. thesis, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

BARBOSA, M., 2009 *Uma abordagem para análise de dados com medidas repetidas utilizando modelos lineares mistos*. Master's thesis, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

CASELLA, G. and R. L. BERGER, 2011 *Inferência Estatística*. Tradução de S.A., Visconde. São Paulo, second edition.

CAVANAUGH, J. E., 1999 A large-sample model selection criterion based on kullback's symmetric divergence. Statistics & Probability Letters **42**: 333–343.

DEMIDENKO, E., 2013 *Mixed models: theory and applications with R*. John Wiley & Sons, Wiley. New York, second edition.

DIGGLE, P. J., P. HEAGERTY, K. Y. LIANG, and S. L. ZEGER, 2002 *Analysis of Longitudinal Data*. Oxford: Oxford University Press, second edition.

EMILIANO, P. C., 2009 *Fundamentos e Aplicações dos critérios de informação: Akaike e Bayesiano*. Master's thesis, Universidade Federal de Lavras, Lavras.

EMILIANO, P. C., 2013 *Critérios de Informação: Como eles se comportam em diferentes modelos?*. Ph.D. thesis, Universidade Federal de Lavras, Lavras.

GILMOUR, A. R., R. THOMPSON, and B. R. CULLIS, 1995 An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51**: 1440–1450.

GURKA, M. J., 2006 Selecting the best linear mixed model under reml. The American Statistician **60**: 19–26.

HARVILLE, D. A., 1977 Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association **358**: 320–338.

KENWARD, M. G. and J. H. ROGER, 1997 Small sample inference for fixed effects from restricted maximum likelihood. Biometrics **53**: 983–997.

KONISHI, S. and G. KITAGAWA, 2008 *Information Criteria and Statistical Modeling*. Springer, New Yourk, first edition.

KULLBACK, S. and R. A. LEIBLER, 1951 On information and sufficiency. The annals of mathematical statistics **22**: 79–86.

LAIRD, N. M. and J. H. WARE, 1982 Randon-effects model for longitudinal data. Biometrics **38**: 963–974.

LIMA, C. G., 1988 *Analise de curvas de crescimento de aves: um enfoque multivariado*. Master's thesis, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

LITTLE, R. J. A. and D. B. RUBIN, 2002 *Statistical Analysis with Missing Data*. Wiley, New York, second edition.

PATTERSON, H. D. and R. THOMPSON, 1971 Recovery of inter-block information when blocks size are unequal. Biometrika **58**: 545–554.

PINHEIRO, J. C., 1994 *Topics in Mixed-Effects Models*. Ph.D. thesis, University of Wisconsin, Madison, WI.

PINHEIRO, J. C. and D. M. BATES, 2000 *Mixed - effects Models in S and S-PLUS*. Springer - Verlang, New York, first edition.

SAS/STAT, 2008 *User's Guide*. SAS Institute Inc, Cary, North Carolina.

SATTERTHWAITE, F. E., 1946 An approximate distribution of estimates of variance components. Biometrics Bulletin **2**: 110–114.

SCHWARZ, G., 1978 Estimating the dimension of a model. The annals of Statistics **6**: 461–464.

SELF, S. G. and K. Y. LIANG, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association **82**: 605–610.

SINGER, J. M., J. S. NOBRE, and M. M. ROCHA, 2015 *Análise de dados longitudinais versão parcial preliminar*. Em Produção.

TEAM, R. C., 2016 *R A language and environment for statistical computing*. <https:www.R-project.org/>, Vienna.

VAIDA, F. and S. BLANCHARD, 2005 Conditional akaike information for mixed-effects models. Biometrika **92**: 351–370.

VERBEKE, G. and G. MOLENBERGHS, 2009 *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media, New York, first edition.

WEST, B. T., K. B. WELCH, and A. T. GATECKI, 2007 *Linear Mixed Models: A Pratical Guide Using Statistical Software*. Chapman & Hall, New York, first edition.

WICKLIN, R., 2013 *Simulating data with SAS*. NC: SAS Institute Inc, Cary, North Carolina.

APÊNCICE

```
# Limpa todos os objetos
rm(list=ls(all=T))
# Carregando pacotes necessários
require(nlme)
require(MASS)
# Repetir o processo 1000 vezes
u=1000
# Vetores para armazenar menores valores de AIC, BIC e KIC
menoraic = c(rep(0,u));menorbic = c(rep(0,u));menorkic = c(rep(0,u))
# Contadores
contaic1 = 0;contaic2=0;contaic3=0;contaic4=0;contaic5=0
contbic1 = 0;contbic2=0;contbic3=0;contbic4=0;contbic5=0
contkic1 = 0;contkic2=0;contkic3=0;contkic4=0;contkic5=0
for(k in 1:u){
# número de repetições para cada indivíduo (tempo)
j=7
# número de indivíduos (frangos)
m=32
# Valores de beta
beta0 = -48.0491; beta1 = 152.98; beta2 = 15.60355; beta5 = 6.1908
# Vetor beta
beta = c(beta0,beta1,beta2,beta5)
# construção da matriz X
tempop = rep(1:7,1,m*j)
tempo2 = tempop^2
col5=c(rep(0,13*j),tempo2[(13*j+1):(m*j)])
X = matrix(cbind(rep(1,m*j),tempop,tempo2,col5),ncol=4)
# construção da matriz Z
Z1=X[,c(2,3)]
Z2=Z1[1:j,]
Z = matrix(rep(0,m*j),nrow=m*j,ncol=m*2)
Z = kronecker(diag(1, m), Z2)
# matriz G (não estruturada)
G = matrix(c(325.50,-46.2661,-46.2661,14.7671),nrow=2,ncol=2)
# vetor de zero
zero = rep(0,nrow(G))
#gerar valores aleatórios para b
b1= mvrnorm(m,zero,G)
b= matrix(0,m*2,ncol=1)
for (i in 1:m){
for (l in 1:2){
if (l==1) b[2*i-1,1]=b1[i,l]
else b[2*i,1]=b1[i,l]
}
}
```

```
# construção da matriz R_i
# Estrutura R_i VC
eps = rnorm(n=nrow(X),mean = 0,sd=sqrt(1450.92) )
#Estrutura R_i AR1
AR=function (n,roh,sigma){
A = matrix(rep(0,n*n),n,n)
for(i in 1:n){
for(j in 1:n){
A[i,j] =roh^{abs(i-j)}}}
C=sigma*A
return(C)
}
sig = AR(7,-0.0329,1450.92)
zero1 = rep(0,7)
eps1= mvrnorm(32, zero1,Sigma = sig)
eps=matrix(rep(0,m*j),ncol=1)
for (r in 1:32){
for (s in 1:7){
eps[7*r-7+s,1]=eps1[r,s]
}
}
#Estrutura R_i ARH1
ARH=function (n,roh,s_1,s_2,s_3,s_4,S_5,S_6,s_7){
A = matrix(rep(0,n*n),n,n)
S = matrix(c(s_1,s_2,s_3,s_4,S_5,S_6,s_7))
for(i in 1:n){
for(j in 1:n){
A[i,j] =S[i,1]*S[j,1]*roh^{abs(i-j)}}}
return(A)
}
sig=ARH(7,0.8640,sqrt(62.7245),sqrt(776.51),sqrt(6847.28),sqrt(7331.15),
sqrt(8528.2),sqrt(21495),sqrt(8311.17))
zero1 = rep(0,7)
eps1= mvrnorm(32, zero1,Sigma = sig)
eps=matrix(rep(0,m*j),ncol=1)
for (r in 1:32){
for (s in 1:7){
eps[7*r-7+s,1]=eps1[r,s]
}
}
# construção de y
y = rep(0,nrow(X))
y = X%*%beta+Z%*%b+eps
#vetor de indivíduos
ind = matrix(0,nrow=j*m,ncol=1)
for(i in 1:m){
ind[((i-1)*j+1):(i*j),1] = rep(i,j)
```

```
}
ind = as.vector(ind)
#fator sexo
sexo<-factor(c(rep(0,13*j),rep(1,m*j-13*j)))
# construção do objeto data.frame
dados <- data.frame(ind,y,tempop,sexo)
# construção do objeto groupedData
dados<-groupedData(y~tempop|ind, data=dados, outer = ~sexo,
labels = list(x="tempo(semana)",y= "Peso corporal (g)"))
############################################################
#modelos candidatos (Exemplo: Seleção das estruturas de covariâncias)
 # 1 ) VC
 par1 = 4 + 3 +1
 ajuste1 = suppressWarnings(lme(fixed=y~tempop+I(tempop^2)+I(tempop^2):sexo,
 random = ~tempop+I(tempop^2)-1,
 data=dados,method = "REML",control= list(returnObject =T))$logLik)
 AIC1 = -2*ajuste1 + 2*par1
 BIC1 = -2*ajuste1 + par1*log(m*j)
 KIC1 = -2*ajuste1 + 3*par1
 # 2) AR(1)
 par2 = 4 + 3 +2
 ajuste2 = suppressWarnings(lme(fixed=y~tempop+I(tempop^2)+I(tempop^2):sexo,
 random = ~tempop+I(tempop^2)-1,
 correlation = corAR1(form = ~tempop|ind),
 data=dados,method = "REML",control= list(returnObject =T))$logLik)
 AIC2 = -2*ajuste2 + 2*par2
 BIC2 = -2*ajuste2 + par2*log(m*j)
 KIC2 = -2*ajuste2 + 3*par2
 # 3) ARH(1)
 par3 = 4+3+8
 ajuste3 = suppressWarnings(lme(fixed=y~tempop+I(tempop^2)+I(tempop^2):sexo,
 random = ~tempop+I(tempop^2)-1,
 correlation = corAR1(form = ~tempop|ind),
 weights = varIdent(form = ~1|tempop),
 data=dados,method = "REML",control= list(returnObject =T))$logLik)
 AIC3 = -2*ajuste3 + 2*par3
 BIC3 = -2*ajuste3 + par3*log(m*j)
 KIC3 = -2*ajuste3 + 3*par3
 # 4) CS
 par5 = 4 +3 +2
 ajuste4 = suppressWarnings(lme(fixed=y~tempop+I(tempop^2)+I(tempop^2):sexo,
 random = ~tempop+I(tempop^2)-1,
 correlation = corCompSymm(form = ~1|ind),
 data=dados,method = "REML",control= list(returnObject =T))$logLik)
 AIC4 = -2*ajuste4 + 2*par4
 BIC4 = -2*ajuste4 + par4*log(m*j)
 KIC4 = -2*ajuste4 + 3*par4
```

```
menoraic[k] = min(AIC1,AIC2,AIC3,AIC4)
menorbic[k] = min(BIC1,BIC2,BIC3,BIC4)
menorkic[k] = min(KIC1,KIC2,KIC3,KIC4)
{if (menoraic[k]==AIC1)
contaic1<-contaic1+1
}
{if (menoraic[k]==AIC2)
contaic2<-contaic2+1
}
{if (menoraic[k]==AIC3)
contaic3<-contaic3+1
}
{if (menoraic[k]==AIC4)
contaic4<-contaic4+1
}
######################BIC###################
{if (menorbic[k]==BIC1)
contbic1<-contbic1+1
}
{if (menorbic[k]==BIC2)
contbic2<-contbic2+1
}
{if (menorbic[k]==BIC3)
contbic3<-contbic3+1
}
{if (menorbic[k]==BIC4)
contbic4<-contbic4+1
}
######################KIC###################
{if (menorkic[k]==KIC1)
contkic1<-contkic1+1
}
{if (menorkic[k]==KIC2)
contkic2<-contkic2+1
}
{if (menorkic[k]==KIC3)
contkic3<-contkic3+1
}
{if (menorkic[k]==KIC4)
contkic4<-contkic4+1
}
}
dados1<-data.frame(contaic1,contaic2,contaic3,contaic4,
contbic1,contbic2,contbic3,contbic4,
contkic1,contkic2,contkic3,contkic4); dados1
```

ANEXOS

Anexo A

Para organização dos dados, o objeto do R comumente usado é o **data.frame**. No entanto para os dados de medidas repetidas, Pinheiro e Bates (2000) complementaram com o objeto **data.frame** fornecendo informações adicionais sobre os dados agrupados. Este objeto é classificado como **groupedData**.

À título de ilustração considere um exemplo de um conjunto de dados que está disponível no pacote *nlme*: `Orthodont` em que foram feitas quatro medidas ortodônticas (aos 8, 10, 12 e 14 anos de idade), em mm, em 16 meninos e 11 meninas.

O código do R

```
require (nlme)
Orthodont
```

indica que primeiramente o pacote *nlme* está sendo carregado e a seguir a saída do programa fornece:

```
Grouped Data: distance ~ age | Subject
distance age Subject    Sex
1       26.0   8     M01   Male
2       25.0  10     M01   Male
3       29.0  12     M01   Male
4       31.0  14     M01   Male
5       21.5   8     M02   Male
6       22.5  10     M02   Male
7       23.0  12     M02   Male
8       26.5  14     M02   Male
.
.
.
61      22.0   8     M16   Male
62      21.5  10     M16   Male
63      23.5  12     M16   Male
64      25.0  14     M16   Male
65      21.0   8     F01 Female
66      20.0  10     F01 Female
67      21.5  12     F01 Female
68      23.0  14     F01 Female
.
.
.
104     19.5  14     F10 Female
105     24.5   8     F11 Female
106     25.0  10     F11 Female
107     28.0  12     F11 Female
108     28.0  14     F11 Female
```

O objeto **groupedData** fornece : a variável reposta (`distance`), a covariável (`age`), o nome do fator (`subject`) que indica qual o indivíduo que está sendo medido ao longo do tempo e o fator (`sex`) que classifica se o indivíduo é do sexo feminino ou masculino.

Para a construção do objeto **groupedData** é necessário que os dados esteja num formato de banco de dados, como por exemplo, a classe **data.frame**. A função que cria um objeto de uma certa classe é chamada de construtor. No caso, a função **groupedData** é o próprio construtor da classe **groupedData** (Pinheiro e Bates, 2000).

Os modelos lineares mistos são ajustados pela função `lme()` por um dos métodos a ser definido pelo usuário: Método da Máxima Verossimilhança (MV) ou o método da Máxima Verossimilhança Restrita (MVR). Se não for especificado pelo usuário, por padrão o modelo será ajustado pelo MVR.

Por padrão a matriz de variância e covariância para os efeitos aleatórios, $\mathbf{G}$, é assumido como uma matriz sem características sistemáticas na função `lme`, ou seja, a matriz $\mathbf{G}$ é assumida como não estruturada. No entanto, muitas vezes é desejável restringir a estrutura da matriz $\mathbf{G}$ caracterizada com poucos parâmetros. O pacote *nlme* fornece a classe **pdMat** que especifica padrões sistemáticos da matriz de variâncias e covariâncias dos efeitos aleatórios (Pinheiro e Bates, 2000).

Tabela anexo1. Padrões da matriz de variâncias e covariâncias dos efeitos aleatórios da classe pdMat

Estrutura	Descrição
pdBlocked	Bloco diagonal
pdCompSymm	Estrutura simetria composta
pdDiag	Estrutura diagonal
pdIdent	Estrutura múltipla de uma identidade
pdSymm	matriz positiva-definida geral

Fonte: Pinheiro e Bates (2000)

Por padrão a matriz $\mathbf{R}_i$ é considerada $\sigma^2\mathbf{I}$, ou seja, a estrutura pdDiag. No entanto, em estruturas de dados com medidas repetidas é muito comum observar que os erros intra-indivíduos são heterocedásticos, ou estes são correlacionados, ou ainda que estes são heterocedásticos e correlacionados.

Segundo Pinheiro e Bates (2000) a função de variância é utilizada para modelar a estrutura de variância dos erros intra-indivíduos utilizando covariáveis. No pacote *nlme* existe uma classe que fornece um conjunto de funções de variância: **varFunc**.

Tabela anexo2. Padrões da estrutura da função de variância da classe varFunc

Estrutura	Descrição
varFixed	variância fixa
varIdent	diferentes variâncias por níveis de um fator
varPower	poder de uma covariável
varExp	exponencial de uma covariável
varComb	combinações de funções de variância

Fonte: Pinheiro e Bates (2000)

Para a utilização das funções de variância na função `lme()`, o argumento utilizado é `weights`.

Para modelar a dependência entre as observações são utilizadas as diferentes estruturas de correlação. No pacote *nlme* existe uma classe de conjuntos de estruturas de correlação: **corStruct**.

Tabela anexo3. Padrões da estrutura de correlação da classe corStruct

Estrutura	Descrição
corCompSymm	simetria composta
corSymm	geral
corAR1	autoregressiva de ordem 1
corCAR1	AR1 com tempo (variável) contínuo
corARMA	ARMA(p,q)
corExp	exponencial
corLin	linear
corGaus	gaussiana

Fonte: Pinheiro e Bates (2000)

Para a utilização das estruturas de correlação na função `lme()`, o argumento utilizado é `correlation`.

Anexo B

Inicialmente os dados serão ajustados utilizando o PROC MIXED e as estimativas dos efeitos fixos e dos parâmetros de covariância serão considerados como sendo "verdadeiras"para o processo de simulação.

```
data frango;
input indiv sex$ peso tempo;
datalines;
1     Femea     122      1
1     Femea     291      2
1     Femea     500      3
1     Femea     712      4
1     Femea    1041      5
1     Femea    1430      6
1     Femea    1760      7
2     Femea     129      1
2     Femea     314      2
.
.
.
31     Macho    2180      7
32     Macho     118      1
32     Macho     277      2
32     Macho     591      3
32     Macho     870      4
32     Macho    1256      5
32     Macho    1738      6
32     Macho    2050      7
;
```

O código a seguir ajusta o peso dos frangos considerando *tempo*, $tempo^2$ e $tempo^2$ interação com *sexo* de efeitos fixos e *tempo* e $tempo^2$ de efeitos aleatórios. A estrutura da matriz **G** é definida

como não estruturada (UN) e a matriz **R** tem a estrutura padrão de componente de variância (VC), que não precisa ser especificada em um comando `repeated`.

```
proc mixed data = frango;
class sex;
model peso = tempo tempo*tempo tempo*tempo*sex / solution outpm=outpm;
random tempo tempo*tempo / subject= indiv type = un;
ods select CovParms SolutionF;
ods output CovParms=CovParms SolutionF=SolutionF;
run;
```

Passo 1: O código a seguir constrói a matriz X e a matriz Z.

```
proc glimmix data=frango outdesign(names novar)=All;
class sex;
model peso = tempo tempo*tempo tempo*tempo*sex;
random tempo tempo*tempo;
ods select ColumnNames;
run;
```

Construir as variáveis para a matriz X e matriz Z.

```
proc iml;
FixedVar = "_X1":"_X5";
RandomVar = "_Z1":"_Z2";
use All;
read all var FixedVar into X;
read all var RandomVar into Z;
close All;
row = nrow(X);
```

Passo 4a : Usar as estimativas dos efeitos fixos obtidos pelo PROC MIXED.

```
use SolutionF; read all var {Estimate} into beta; close;
eta = X*beta; /* preditor linear */
print X beta eta;
```

Passo 2: Utiliza as estimativas dos parâmetros de covariância obtida pelo PROC MIXED e define e matriz **G** como sendo não estruturada.

```
use CovParms; read all var {Estimate} into var; close;
print var;
varT = var[1]; /* Var(tempo) */
covTT = var[2]; /* Var(tempo) */
varTT = var[3]; /* Var(tempo*tempo) */
sigma2 = var[4]; /* sigma2 = Var(Error) */
v = { 325.50492  -46.26615  14.7671};
G = sqrvech(v);
print G ;
```

Passo 3 e 4:

```
call randseed(1);
zero = repeat(0, nrow(G));
NumSamples = 7;
gamma = RandNormal(NumSamples, zero, G);
eps = J(224, NumSamples);
call randgen(eps, "Normal", 0, sqrt(1611.5015));
Y = eta + Z*gamma` + eps;
```