SIMONE INOE ARAÚJO

USO DE MODELOS DE REGRESSÃO ALEATÓRIA NA ANÁLISE DE DADOS LONGITUDINAIS NO MELHORAMENTO GENÉTICO VEGETAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de "Doctor Scientiae".

VIÇOSA
MINAS GERAIS - BRASIL
2005

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
A662u
2005

Araújo, Simone Inoe, 1975-
Uso de modelos de regressão aleatória na análise de dados longitudinais no melhoramento genético vegetal / Simone Inoe Araújo. – Viçosa : UFV, 2005.
x, 111f : il. ; 29cm.

Orientador: Adair José Regazzi.
Tese (doutorado) - Universidade Federal de Viçosa.

Referências bibliográficas: f.107-109

1. Plantas - Melhoramento genético – Simulação por computador. 2. Plantas - Melhoramento genético - Métodos estatísticos. 3. Análise de variância. 4. Análise de regressão. 5. Máxima verossimilhança restrita. I. Universidade Federal de Viçosa. II. Título.

CDD 22.ed. 631.52

SIMONE INOE ARAÚJO

USO DE MODELOS DE REGRESSÃO ALEATÓRIA NA ANÁLISE DE DADOS LONGITUDINAIS NO MELHORAMENTO GENÉTICO VEGETAL

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de "Doctor Scientiae".

APROVADA: 20 de maio de 2005.

Prof. Cosme Damião Cruz
(Conselheiro

Prof. José Marcelo Soriano Viana
(Conselheiro)

Prof. Carlos Henrique Osório Silva

Prof. Cláudio Vieira de Araújo

Prof. Adair José Regazzi
(Orientador)

Aos meus pais, Paulo Inoe (*in memoriam*) e Sumiko Inoe,

DEDICO

Ao meu esposo Cláudio Vieira de Araújo e aos meus filhos João Pedro, Guilherme e Mariana,

OFEREÇO

AGRADECIMENTOS

A Deus, por tornar tudo possível.

À Universidade Federal de Viçosa e ao curso de Pós-Graduação em Genética e Melhoramento, pela oportunidade de realização do curso.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela concessão da bolsa de estudos.

Ao professor Adair José Regazzi, pelos ensinamentos, pela orientação, pela amizade e pelo apoio.

Ao meu esposo Cláudio Vieira de Araújo, pelos ensinamentos, auxílios e pela compreensão durante a realização deste trabalho.

Aos professores Paulo Sávio Lopes e Robledo de Almeida Torres, pelo apoio.

Ao professor José Marcelo Soriano Viana pelos ensinamentos, pelo aconselhamento e pelas sugestões.

Ao professor Cosme Damião Cruz, pelos ensinamentos, pelo incentivo e pelo aconselhamento.

Aos professores Carlos Henrique Osório Silva e Renato Ribeiro de Lima, pelas críticas e sugestões.

Aos amigos e colegas de pós-graduação, Alex, Andréa Barcelos, Amauri, Aurinelza Teixeira, Antônio Policarpo, Edivânia Evangelista, Eduardo Cruz, Elizângela, Felipe, Guilherme, Giselle, João Francisco, Jaime, Jane, José Elivalto, José Marques, Lindenberg, Paulo Carneiro, Kécya, Rachel, Ricardo e William, pela amizade e convívio.

Aos amigos Rodolpho de Almeida Torres e Carla; Aldrin Vieira Pires e Ivy pela amizade.

À amiga Luciara Celi Chaves e seus familiares, pela amizade e pelo apoio.

Às funcionárias da secretaria do curso de Pós-Graduação em Genética e Melhoramento, Conceição e Rita, pelos serviços prestados e pela amizade.

Aos demais professores, funcionários e alunos do curso de Pós-graduação em Genética e Melhoramento, pelo companheirismo e pela amizade.

Aos amigos Jane,Ademir e Milena, pelo convívio e grande amizade.

A minha família, pela base e pela formação.

Aos meus irmãos Cláudia, Marcelo e Ana Paula, pelo constante apoio e por estarem sempre torcendo por mim.

Aos meus filhos João Pedro e Guilherme, por me proporcionarem a maior alegria da minha vida: ser mãe.

A todos que, de algum modo, contribuíram para a realização deste trabalho.

CONTEÚDO

RESUMO

ARAÚJO, Simone Inoe, D.S., Universidade Federal de Viçosa, maio de 2005. **Uso de Modelos de Regressão Aleatória na análise de dados longitudinais no melhoramento genético vegetal**. Orientador: Adair José Regazzi. Conselheiros: Cosme Damião Cruz e José Marcelo Soriano Viana.

Os objetivos deste estudo foram analisar, via simulação de dados, o efeito de diferentes pressuposições quanto à variância dos efeitos ambientais, frente a dados com determinada estrutura de variâncias, e verificar o comportamento de diferentes estratégias de análise frente ao desbalanceamento dos dados. Foram simulados dados referentes a um teste de progênie, do cruzamento de 30 progenitores masculinos com três genitores femininos diferentes cada um, onde cada cruzamento deu origem a dez indivíduos, distribuídos em três locais diferentes. O efeito fixo de local foi gerado de forma a não apresentar diferenças estatísticas significativas. Para cada indivíduo da prole foram geradas informações fenotípicas em cinco idades diferentes. Portanto, o arquivo de dados consistiu de 1020 indivíduos no total, sendo que 900 indivíduos apresentaram registros nas cinco idades, totalizando 4500 registros de produção. Para estudar o efeito da heterogeneidade das variâncias ambientais, em modelos de regressão aleatória adotou-se, modelos que ajustaram uma função polinomial de segundo grau para o efeito genético aditivo e de ambiente permanente e que ajustaram uma função polinomial apenas para o efeito genético aditivo foram analisados, considerando-se ou não a heterogeneidade da variância do efeito de ambiente temporário, gerando-se assim, quatro diferentes modelos de regressão aleatória. Além disso, os modelos de regressão aleatória, repetibilidade e multi-característica foram avaliados sob diferentes níveis de desbalanceamento dos dados. O modelo de regressão aleatória mais adequado foi aquele que considerou a heterogeneidade de variâncias dos efeitos de ambiente permanente e temporário. Assumir pressuposições incorretas sobre a estrutura de covariância dos efeitos

aleatórios do modelo conduziu à alterações nas estimativas de componentes de covariância e nas estimativas dos parâmetros genéticos. Sob desbalanceamento sem seleção, todos os modelos apresentaram estimativas de herdabilidade bastante semelhantes aos resultados obtidos quando se considerou o conjunto de dados completos. Entretanto, quando se considerou o efeito da seleção, modelos de regressão aleatória com até 10% de desbalanceamento não promoveram alterações nas estimativas de componentes de variância.

ABSTRACT

ARAÚJO, Simone Inoe, D.S., Universidade Federal de Viçosa, may 2005. **Use of random regression models in longitudinal analysis data in genetic plant breeding**. Adviser: Adair José Regazzi. Commitee members: Cosme Damião Cruz and José Marcelo Soriano Viana.

The aim of this study was to analyze the effect of assuming different assumptions about environmental variance effects to data with certain variance structure, and verify genetic parameters estimates in different analysis strategies behind unbalanced data. A progeny test data was simulated, by crossing 30 male with three different female, where each crossing originated ten individuals, distributed in three different places. The fixed effect of place was generated in order not to present significant statistical difference. For each individual offspring, phenotypic information were generated in five different ages. Then, the data consisted in a total of 1020 individuals, in what 900 of them presented information in the five ages, computing 4500 observations of production. To verify the importance of consider or not the environmental heterogeneity of variance, in random regression models, models that adjusted a second polynomial function both for the additive genetic as for the permanent environmental effect and adjusted a polynomial function only for the additive genetic effect were analyzed, considering or not the variance heterogeneity of the temporary environmental effect, then generating four different random regression models. Moreover, the single-trait random regression model, the repeatability model and the multiple-trait model were analyzed on different lost of information levels. The most adequate random regression model was the one who considered both, the variance heterogeneity of permanent environmental effect, and the temporary environmental effect. Assuming the incorrect assumptions about the covariance structure of random effects of the model, conducted to change in the covariance components estimates and in the genetic

parameters estimates. With incomplete data, without selection, all the models presented heritability estimates very similar to the results when complete data were considered. However, when effect of selection was considered, random regression models with less or equal to 10% of lost of information didn’t conduct to change in the covariance components estimates.

INTRODUÇÃO GERAL

A realização de medidas no tempo ou no espaço, em um mesmo indivíduo ou em mesma progênie, é prática comum no melhoramento de espécies perenes, como a erva-mate, a seringueira, o cacaueiro, o coqueiro, o cupuaçuzeiro, o guaranazeiro e em espécies florestais, em medidas incrementais de crescimento. Como as características de interesse no melhoramento de plantas perenes se expressam mais de uma vez em um mesmo indivíduo, geram dados longitudinais, sendo essas características denominadas de dimensão infinita (RESENDE, 2002).

Segundo DIAS e RESENDE (2001), uma das características que diferenciam o melhoramento de espécies perenes arbóreas, como o cacaueiro, por exemplo, das demais, é que o melhoramento dessas espécies exige a utilização de métodos de seleção mais acurados, que necessitam avaliações sucessivas no tempo, feitas na mesma unidade de seleção. Essas espécies perenes apresentam aspectos fisiológicos, biológicos, reprodutivos e fitotécnicos muito peculiares, destacando-se ciclos produtivo e reprodutivo longo, a sobreposição de gerações, e a expressão fenotípica de diversos caracteres ao longo do tempo, com mudanças temporais no controle genético deles.

Dentre as principais contribuições da Genética Biométrica para o melhoramento de plantas perenes, destaca-se a possibilidade da predição dos valores genéticos aditivos (com vistas à propagação sexuada) e valores genotípicos (com vistas à propagação vegetativa) de todos os indivíduos candidatos à seleção (DIAS e RESENDE, 2001)

Os modelos de regressão aleatória foram desenvolvidos para analisar dados que são obtidos por sucessivas medições da característica durante a vida de um indivíduo, gerando as denominadas medidas repetidas. No melhoramento vegetal ainda são poucos os trabalhos utilizando essa metodologia de análise.

O modelo de regressão aleatória foi utilizado pela primeira vez, para avaliação genética em melhoramento genético animal, por SCHAEFFER e

DEKKERS (1994). Esses autores sugeriram esse modelo no melhoramento genético de gado de leite, para descrever a curva de lactação de registros do dia do controle (test-day). Tal procedimento, permitiu que a forma da curva padrão de lactação fosse diferente para cada animal, por meio da inclusão de um coeficiente de regressão aleatória para cada animal. A forma da curva de lactação para um determinado animal poderia ser descrita como dois grupos de regressão sobre o número de dias em produção. Um conjunto de coeficientes de regressão fixa para todos os animais pertencentes à mesma subclasse de idade-estação de parto, descrevendo a forma geral da curva para todos os animais daquela subclasse e um conjunto de coeficientes de regressão aleatória para cada animal individualmente, descrevendo os desvios em relação à curva descrita pela regressão fixa. Isso permitiu que cada animal apresentasse uma curva de lactação própria.

Devido à pouca utilização da técnica de regressão aleatória no melhoramento vegetal de culturas perenes, é preciso definir pontos, que no melhoramento animal já têm sido bastante elucidados. Este trabalho teve por objetivos avaliar modelos com diferentes pressuposições, com relação às variâncias ambientais e avaliar o efeito da perda de informações em diferentes estratégias de análise de dados longitudinais para sua utilização no melhoramento de espécies vegetais perenes.

REVISÃO DE LITERATURA

Um conjunto de observações provenientes de várias medições, tomadas de forma seqüencial, na mesma unidade experimental ao longo do tempo ou espaço, recebe a denominação de dados longitudinais ou medidas repetidas. De acordo com VAN DER WERF e SCHAEFFER (1997), características assim merecem um tratamento estatístico especial, pois pode existir uma estrutura de covariâncias entre as medidas repetidas, e para se inferir corretamente utilizando-se dados com medidas repetidas, pode ser importante modelar esta estrutura de covariâncias.

Segundo RESENDE (2002), a realização de medidas no tempo ou no espaço, em um caráter no indivíduo é prática comum no melhoramento de espécies perenes, como a erva-mate, a seringueira, o cacaueiro, o coqueiro, o cupuaçuzeiro, o guaranazeiro e em espécies florestais, em medidas incrementais de crescimento. Como as características de interesse no melhoramento de plantas perenes se expressam mais de uma vez em um mesmo indivíduo, geram dados longitudinais, sendo essas características denominadas de dimensão infinita. Entretanto, nos últimos tempos, este tipo de dado tem recebido atenção especial, principalmente por melhoristas da área animal. A produção de leite nos diversos controles leiteiros é um exemplo típico de medidas repetidas, que juntamente com a análise de crescimento de animais, são os mais trabalhados no contexto do melhoramento animal. No melhoramento vegetal de espécies arbóreas perenes, como o eucalipto, as diversas medições do diâmetro à altura do peito (DAP) representam um exemplo de um conjunto de dados longitudinais.

Dentre as alternativas de análise de dados longitudinais estão as análises por meio de modelos de repetibilidade e de modelos multi-características. Nos modelos de repetibilidade é pressuposto que a característica é a mesma ao longo das várias medidas no tempo. Sendo assim, assume-se que todas as medidas apresentam correlação igual à unidade, e, portanto, todas as (co)variâncias genéticas e fenotípicas entre as diferentes medidas são de mesma magnitude. Entretanto, RESENDE (2002) relatou que nem todas as culturas atendem a essa pressuposição e por isso, em espécies florestais, este modelo é

raramente aplicado. Normalmente, para características de crescimento a correlação entre medidas tomadas ao longo do tempo diminui quando o espaço de tempo entre elas aumenta. No modelo multi-característica uma escala contínua é dividida arbitrariamente em intervalos, alguns pontos da curva de crescimento são amostrados, e os indivíduos são selecionados, não sendo possível fazer inferências em pontos intermediários da curva. Neste modelo as medidas realizadas no tempo são consideradas como características distintas e correlacionadas. Embora este seja o modelo mais correto, quando se têm várias medidas, ele utiliza um elevado número de parâmetros, inviabilizando o seu uso em face às limitações computacionais e também com relação à interpretação desses parâmetros. De acordo com VAN DER WERF e SCHAEFFER (1997), tal abordagem apresenta desvantagens. A primeira é que é ajustada uma estrutura de covariância descontínua, onde realmente é contínua. Outra desvantagem de um modelo multi-característica, para características em muitas idades diferentes, é que a matriz de correlação é não estruturada.

Ao invés da aplicação de modelos com um número finito de características, uma abordagem de dimensão infinita (trajetória) pode ser considerada.

MODELOS DE REGRESSÃO ALEATÓRIA

Os modelos de regressão aleatória foram desenvolvidos para analisar dados que são obtidos por sucessivas medições de uma característica durante a vida de um indivíduo, gerando as denominadas medidas repetidas.

O modelo de regressão aleatória foi utilizado pela primeira vez, para avaliação genética em melhoramento animal, por SCHAEFFER e DEKKERS (1994). Esses autores sugeriram esse modelo no melhoramento genético de gado de leite, para descrever a curva de lactação de registros do dia do controle (test-day). Tal procedimento permitiu que a forma da curva padrão de lactação fosse diferente para cada animal, por meio da inclusão de um coeficiente de regressão aleatória para cada animal. A forma da curva de lactação para um determinado animal poderia ser descrita como dois grupos de regressão sobre o

número de dias em produção. Um conjunto de coeficientes de regressão fixo para todos os animais pertencentes à mesma subclasse de idade-estação de parto, descreve a forma geral da curva para todos os animais daquela subclasse, e um conjunto de coeficientes de regressão aleatória para cada animal individualmente, descreve os desvios em relação à curva descrita pela regressão fixa. Isso permitiu que cada animal apresentasse uma curva de lactação própria. Com o desenvolvimento dessa metodologia de análise para dados de produção de leite, foram surgindo diversos trabalhos nesta linha de pesquisa, tais como JAMROZIK e SHAEFFER (1997), JAMROZIK et al. (1997), OLORI et al. (1999), VAN der WERF et al. (1998), REKAYA et al. (1999), BROTHERSTONE et al. (2000), e MEYER (2004).

A aplicação da teoria de modelos de regressão aleatória para a avaliação genética em gado de leite, usando registros da produção no dia do controle, é a mais conhecida. Posteriormente surgiram outras aplicações, dentre as quais pode-se citar a sua aplicação a características de crescimento em todas as espécies, interações genótipo x ambiente, a dados de análise de sobrevivência, etc. Os modelos de regressão aleatória permitem ao pesquisador estudar mudanças na variabilidade genética com o tempo, e também, a seleção de indivíduos para alterar o padrão de resposta sobre o tempo (SCHAEFFER, 2004).

No Brasil, modelos de regressão aleatória têm sido bastante utilizados por pesquisadores da área de melhoramento animal, tais como SAKAGUTI (2000) e ALBUQUERQUE (2003), em crescimento de bovinos de corte, e EL FARO (2002), COSTA et al. (2002), COBUCCI (2002) e ARAÚJO (2003), a dados de produção de leite.

Segundo SCHAEFFER (2004), modelos de regressão aleatória poderiam ser perfeitamente aplicados ao crescimento de plantas, tais como culturas que crescem rapidamente ou árvores que crescem lentamente. No melhoramento vegetal, modelos de regressão aleatória foram utilizados para a avaliação genética de 45 progênies de *Eucalyptus urophylla*, por RESENDE et al.(2001), para a característica diâmetro à altura do peito. O objetivo foi comparar a utilização de modelos de regressão aleatória, o modelo univariado para cada idade separadamente, e o modelo de repetibilidade. Os

autores concluíram que os modelos de regressão aleatória na análise genética conduziram a melhores resultados do que os modelos univariados e de repetibilidade. Em modelos de regressão aleatória no contexto de modelos mistos, cada indivíduo apresentará uma curva própria que permite predizer valores genéticos associados à qualquer idade dentro do intervalo estudado. É possível predizer valores genéticos de indivíduos avaliados em diferentes idades e com diferentes números de idades mensuradas e obter a projeção desses valores genéticos para uma idade padrão, objetivando ter uma importante ferramenta para efeito de ordenamento e seleção dos indivíduos.

A teoria sobre modelos de regressão aleatória (MRA) foi inicialmente proposta por HENDERSON JÚNIOR (1982). De acordo com o autor, se há um coeficiente de regressão pertencente a cada indivíduo em um experimento, e se há uma amostra aleatória de indivíduos, então os coeficientes de regressão devem ser considerados como aleatórios. Por exemplo, se a produção de leite está sendo examinada como uma função do estágio de lactação, entre outras variáveis, deveria existir uma regressão geral, regressões individuais, devido a diferenças genéticas e outras diferenças individuais entre animais, não explicadas pelo estágio de lactação.

SCHAEFFER (2000) apresentou uma formulação sobre a teoria de modelos de regressão aleatória, onde considerou classes de estatura de vacas da raça holandesa, mensuradas em 7 diferentes épocas durante a vida do animal. Segundo o autor, dado que a relação entre as classes de estatura e a idade é curvilínea, o seguinte modelo poderia ser ajustado:

$$\mathbf{y} = a_0 + a_1(\mathbf{A}) + a_2(\mathbf{A})^2 + e,$$

onde **A** representa a idade durante a classificação (em meses) e o $\mathbf{R}^2$ para o ajuste desse modelo foi de 0,98.

As diferenças entre os animais em suas taxas de maturidade, mensuradas pelas avaliações de estatura, poderiam ser incluídas na equação acima, de modo que a **k**-ésima observação sobre o **j**-ésimo animal poderia ser descrita como:

$$\mathbf{y}_{jk} = a_{oj} + a_{1j}(\mathbf{A}) + a_{2j}(\mathbf{A})^2 + e_{jk},$$

que pode ser expandida, escrevendo um modelo para cada a_{ij}, como a seguir:

$$a_{ij} = \mu_i + \mathbf{a}_{ij} + \mathbf{p}_{ij} + \mathbf{e}_{ij},$$

em que μ_i é o efeito médio geral; $\mathbf{a}_{ij}$ é um efeito aleatório genético aditivo; $\mathbf{p}_{ij}$ é um efeito aleatório de ambiente permanente e $\mathbf{e}_{ij}$ é um efeito aleatório residual.

Substituindo o sub-modelo anterior para cada a_{ij} na equação de regressão que relaciona a estatura com a idade, teremos:

$$\begin{aligned}\mathbf{y}_{jk} &= (\mu_0 + \mathbf{a}_{0j} + \mathbf{p}_{0j} + \mathbf{e}_{0j}) \\ &+ (\mu_1 + \mathbf{a}_{1j} + \mathbf{p}_{1j} + \mathbf{e}_{1j})(\mathbf{A}) \\ &+ (\mu_2 + \mathbf{a}_{2j} + \mathbf{p}_{2j} + \mathbf{e}_{2j})(\mathbf{A})^2 \\ &+ e_{jk}\end{aligned}$$

Assim,

$$\begin{aligned}\mathbf{y}_{jk} &= \mu_0 + \mu_1(\mathbf{A}) + \mu_2(\mathbf{A})^2 \\ &+ \mathbf{a}_{0j} + \mathbf{a}_{1j}(\mathbf{A}) + \mathbf{a}_{2j}(\mathbf{A})^2 \\ &+ \mathbf{p}_{0j} + \mathbf{p}_{1j}(\mathbf{A}) + \mathbf{p}_{2j}(\mathbf{A})^2 \\ &+ \mathbf{e}_{0j} + \mathbf{e}_{1j}(\mathbf{A}) + \mathbf{e}_{2j}(\mathbf{A})^2 \\ &+ e_{jk}\end{aligned}$$

Portanto, esta equação representa:

$y =$ regressões fixas+

regressões aleatórias de efeitos genéticos de animal+

regressões aleatórias de efeitos permanentes de animal+

regressões aleatórias de efeitos residuais+

um termo para o efeito do erro aleatório

Portanto, o modelo de regressão aleatória abriga dois grupos de regressão. O primeiro grupo está relacionado com a regressão fixa para todos os indivíduos, pertencentes à mesma classe de efeito fixo (por exemplo, local) e descreve a forma geral da trajetória. O segundo grupo está relacionado com a regressão aleatória para cada indivíduo e descreve os desvios em relação à regressão fixa, permitindo que, para cada

indivíduo seja predita, a sua própria trajetória. Vale ressaltar que no modelo acima, considerou-se que as variâncias do efeito de ambiente temporário não foram constantes para os indivíduos.

ALBUQUERQUE (2003) apresentou a estrutura de um modelo de regressão aleatória, usando polinômios de Legendre da idade (tempo) como covariável independente, ou seja,

$$\mathbf{y}_{ij} = \mathbf{F} + \sum_{m=0}^{k_a - 1} a_{im} f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{k_q - 1} d_{im} f_m(\mathbf{a}^*_{ij}) + e_{ij} ,$$

onde $\mathbf{y}_{ij}$ é a medida do indivíduo **i** no tempo **j**, **F** é um conjunto de efeitos fixos pertencentes a $\mathbf{y}_{ij}$(incluindo sempre uma regressão fixa sobre polinômios do tempo), a_{im} e d_{im} são os **m**-ésimos coeficientes de regressão aleatória genético aditivo e de ambiente permanente para o indivíduo **i**, respectivamente, e $\mathbf{k}_a$ e $\mathbf{k}_q$ são as ordens de ajuste dos polinômios correspondentes.

Em notação matricial, este modelo pode ser escrito como:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zp} + e ,$$

em que **b** representa o vetor de efeitos fixos do modelo, **a** inclui a_{im}, **p** inclui d_{im}, e **X** e **Z** são as matrizes de incidência dos efeitos fixos e de covariáveis dos efeitos aleatórios, respectivamente. A pressuposição do modelo é dada por:

$\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} \sim \mathbf{N}(Æ, \mathbf{V})$, em que Æ é um vetor de zeros e $\mathbf{V} = \begin{bmatrix} \mathbf{G} Ä \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} Ä \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_e \end{bmatrix}$.

G é uma matriz de variâncias e covariâncias genéticas dos coeficientes de regressão aleatória, considerados idênticos em todos os indivíduos, **A** é a matriz dos coeficientes de parentesco entre os indivíduos, **P** é a matriz de variâncias e covariâncias de ambiente permanente dos coeficientes de regressão aleatória, considerados idênticos em todos os indivíduos, **I** é uma matriz identidade, σ^2_e é a variância residual e Ä é o operador produto direto. É possível ajustar modelos de regressão aleatória de modo que o efeito de ambiente permanente seja aleatório ou constante ao longo das idades. Entretanto,

COBUCCI (2002) verificou que a inclusão dos coeficientes de regressão aleatória para descrever o efeito de ambiente permanente promove uma definição mais precisa dos efeitos genéticos e não-genéticos que influenciam a produção. JAMROZIK e SHAEFFER (1997) estudaram modelos de regressão aleatória, onde consideraram constante o efeito de ambiente permanente durante a lactação, e evidenciaram a necessidade de ajustar uma função de regressões aleatórias. Os autores consideraram o efeito residual não constante para grupos de dias em lactação e verificaram que essas variâncias seguem um determinado padrão, ou seja, são maiores no início e no final da lactação. COBUCCI (2002) comentou que, além da inclusão dos coeficientes de regressão aleatória para o efeito de ambiente permanente, deve-se averiguar a utilização da variância residual heterogênea. Entretanto, o que dificulta a adoção de tais efeitos como sendo variáveis é a limitação computacional.

Ao se ajustar um modelo de regressão aleatória, fica implícita a idéia dos dois grupos de regressão anteriormente citados.

FUNÇÕES DE COVARIÂNCIA

O uso de Funções de Covariância

KIRKPATRICK e HECKMAN (1989) definiram como caracteres de dimensão infinita, ou seja, caracteres oriundos de dados longitudinais, aqueles onde o fenótipo de um indivíduo é descrito por uma função, em vez de um número finito de mensurações. Como exemplo de caracteres de dimensão infinita, os autores relataram a trajetória de crescimento de um indivíduo, que pode ser considerada como uma função que relaciona a idade com algumas mensurações do tamanho corporal. Assim, o tamanho do indivíduo para cada idade diferente pode ser considerado uma característica distinta, havendo, portanto, um número infinito de idades. VAN DER WERF e SCHAEFFER (1997) e TIJANI et al. (1998) definiram a função de covariância como uma função contínua, que

fornece as variâncias e covariâncias de características mensuradas em diferentes pontos em uma trajetória.

Em uma abordagem de dimensão infinita, a estrutura de covariância pode ser modelada como uma função de covariância entre as épocas $\mathbf{t_i}$ e $\mathbf{t_j}$, por exemplo, que representam pontos ao longo de uma trajetória definida. Sendo assim, a função de covariância permite analisar uma mudança gradual de covariância no tempo, além de estimar variâncias nos pontos e covariâncias entre todos os pares de pontos ao longo de uma trajetória, ainda que não se tenha ou se tenham poucas observações feitas para esses pontos, porém, utilizando a informação de todas as outras medidas.

No melhoramento genético, o interesse em características que mudam com o tempo está nos parâmetros genéticos, que podem fornecer a informação de como trabalhar, geneticamente, tais alterações. KIRKPATRICK e HECKMAN (1989) comentaram que os modelos de função de covariância apresentam vantagens sobre os modelos multi-características em dados longitudinais, tais como: modelagem com acurácia da estrutura de covariância dos tratamentos descritos, são capazes de predizer estruturas de covariância em qualquer ponto ao longo de uma escala contínua (tempo), permitindo que as covariâncias entre idades que não tenham sido mensuradas sejam facilmente obtidas por interpolação e permitem analisar padrões de covariância que são associados com mudanças particulares da característica ao longo da trajetória, por meio de um conjunto de autovalores e autofunções.

VAN DER WERF e SCHAEFFER (1997) evidenciaram que as funções de covariância permitem modelar a estrutura de covariâncias das características de forma acurada, podem predizer a estrutura de covariâncias em qualquer ponto em uma escala contínua no tempo, permitem a utilização de medidas tomadas em qualquer ponto ao longo de uma trajetória, sem a necessidade de correção para uma idade padrão e permitem analisar padrões de covariância associados a alterações da característica ao longo da trajetória.

KIRKPATRICK et al. (1990), desenvolveram dois métodos de estimação de funções de covariância aditiva. O primeiro método, denominado de método de ajuste

completo, procura ajustar um polinômio ortogonal de grau **k**, utilizando **k** funções polinomiais, sendo **k=t**, o número de idades mensuradas (**t**). No segundo método, denominado de método de ajuste reduzido, o grau do polinômio é menor do que o número de idades mensuradas (**k < t**). Ambos os casos iniciam-se com a obtenção de uma matriz de covariância genética aditiva, referente a mensurações em **n** idades ($\hat{\mathbf{G}}_{\mathbf{nxn}}$). Várias técnicas podem ser utilizadas para estimar uma função de covariância. Os autores justificaram a escolha de uma família de métodos que envolvem ajustes de funções ortogonais para os dados, pelo fato de que, funções ortogonais são freqüentemente utilizadas em análises dos padrões de variância genética em trajetórias de crescimento. Sendo assim, um par de funções f_i e f_j são ditas normais e ortogonais sobre um intervalo [**u**,**v**] se:

$$\int_{u}^{v} f_i(x) f_j(x) dx = 0 \quad \text{e} \quad \int_{u}^{v} f_i^2(x) dx = 1 .$$

Tanto no método de ajuste reduzido, como no método de ajuste completo, são utilizados os polinômios de Legendre. Nestes polinômios o intervalo [**u**, **v**] é definido como sendo de -1 a 1. A função de covariância genética aditiva entre as mensurações nas idades $\mathbf{a_1}$ e $\mathbf{a_2}$, é linearmente representada pela expressão:

$$\mathbf{f}(\mathbf{a}_1^*, \mathbf{a}_2^*) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} [\mathbf{K}]_{ij} f_i(\mathbf{a}_1^*) f_j(\mathbf{a}_2^*)$$

em que **K** é substituída por seu estimador $\hat{\mathbf{K}}$, que é uma matriz de coeficientes associada à função de covariância, com elementos constantes e que dependem da função de covariância e da família de polinômios ortogonais (f), que está sendo utilizada. $\mathbf{a_i}^*$ é a idade mensurada e deve ser padronizada, por meio da expressão:

$$\mathbf{a}_i^* = \mathbf{u} + \frac{\mathbf{v} - \mathbf{u}}{\mathbf{a}_{max} - \mathbf{a}_{min}} (\mathbf{a}_i - \mathbf{a}_{min}) = -\mathbf{1} + \frac{\mathbf{2}}{\mathbf{a}_{max} - \mathbf{a}_{min}} (\mathbf{a}_i - \mathbf{a}_{min})$$

em que $\mathbf{a}_{max}$ e $\mathbf{a}_{min}$ são as maiores e menores idades mensuradas, respectivamente.

No método de ajuste completo (**k=t**), a função de covariância pode ser escrita em notação matricial por:

$$\hat{\mathbf{G}} = f\hat{\mathbf{K}}f'$$

em que $\hat{\mathbf{G}}$ é a matriz de covariâncias genéticas de ordem **t** (**t** valores genéticos) por **t** (idades mensuradas), f é uma matriz de polinômios ortogonais **t** x **k** (**k** é o grau do polinômio). A matriz f pode ser escrita como **M**L, sendo **M** uma matriz **t** x **k** com suas colunas correspondentes às idades padronizadas e L é uma matriz de ordem **k**, referente aos coeficientes dos polinômios ortogonais. A equação anterior pode ser reescrita na forma:

$$\hat{\mathbf{K}} = f^{-1}\hat{\mathbf{G}}(f^{-1})' .$$

A matriz de coeficientes da função de covariância (**C**) pode ser então obtida por meio da seguinte expressão:

$$\mathbf{C} = L\hat{\mathbf{K}}L' .$$

A seguir, será ilustrada a obtenção de uma função de covariância, para descrever a estrutura de covariâncias para medidas obtidas do peso corporal de indivíduos com 2, 3, 4 e 5 semanas de idade. O vetor de idades padronizadas é: **a** = [-1; -0,33; 0,33; 1]. Uma função de covariância pode ser estimada com base na matriz de variância e covariância aditiva previamente estimada e exibida como:

$$\hat{\mathbf{G}} = \begin{bmatrix} \mathbf{1,50} & \mathbf{1,54} & \mathbf{1,59} & \mathbf{1,38} \\ \mathbf{1,54} & \mathbf{2,20} & \mathbf{2,17} & \mathbf{2,04} \\ \mathbf{1,59} & \mathbf{2,17} & \mathbf{2,90} & \mathbf{2,68} \\ \mathbf{1,38} & \mathbf{2,04} & \mathbf{2,68} & \mathbf{3,20} \end{bmatrix}$$

O **k**-ésimo polinômio de Legendre ortogonal e normalizado (que compõe a matriz de polinômios f) podem ser obtidos por meio da seguinte expressão:

$$\mathbf{P_{n+1}(x) = \frac{1}{n+1}\left[(2n+1)x\,P_n(x) - n\,P_{n-1}(x)\right].}$$

Essa quantidade pode ser normalizada por:

$$\mathbf{f_n(x) = \sqrt{\frac{2n+1}{2}}P_n(x)}\,.$$

Inicialmente, definimos $\mathbf{P_0(x) = 1}$ e $\mathbf{P_1(x) = x}$. Assim, o cálculo do polinômio de ordem 1 é obtido por:

$$\mathbf{f_0(x) = \sqrt{\frac{1}{2}}P_0(x) = \sqrt{\frac{1}{2}}(1) = 0.7071}\,.$$

O polinômio de ordem 2 pode ser obtido por:

$$\mathbf{f_1(x) = \sqrt{\frac{3}{2}}P_1(x) = \sqrt{\frac{3}{2}}x = 1.2247x}\,.$$

O polinômio de ordem 3 é obtido por:

$$\mathbf{P_{1+1}(x) = P_2(x) = \frac{1}{2}\left[3x\,P_1(x) - P_0(x)\right]}$$

$$\mathbf{P_2(x) = \frac{1}{2}(3x.x - 1)}$$

$$\mathbf{P_2(x) = \frac{1}{2}(3x^2 - 1)}$$

$$\mathbf{P_2(x) = \frac{3x^2}{2} - \frac{1}{2}},$$

então,

$$\mathbf{f_2(x) = \sqrt{\frac{5}{2}}\left(\frac{3x^2}{2} - \frac{1}{2}\right) = -0.7906 + 2.3717x^2}.$$

O polinômio de ordem 4 é obtido por:

$$\mathbf{P_{2+1}(x) = P_3(x) = \frac{1}{3}\left[5x\left(\frac{3x^2}{2} - \frac{1}{2}\right) - 2.x\right]}$$

$$\mathbf{P_3(x) = \frac{1}{3}\left[\frac{15x^3}{2} - \frac{9x}{2}\right]}$$

$$\mathbf{P_3(x) = \frac{5x^3}{2} - \frac{3x}{2}}$$

$$f_3(\mathbf{x}) = \sqrt{\frac{7}{2}}\left(\frac{5x^3}{2} - \frac{3x}{2}\right) = \mathbf{-2.8062 + 4.6771x^3}$$

A matriz (**C**), com coeficientes da função de covariância pode ser estimada com os quatro primeiros polinômios (um ajuste de terceiro grau). Na forma matricial, a matriz L para os quatro primeiros polinômios de Legendre é uma matriz 4 x 4:

$$L = \begin{bmatrix} \mathbf{0,7071} & \mathbf{0} & \mathbf{-0,7906} & \mathbf{0} \\ \mathbf{0} & \mathbf{1,2247} & \mathbf{0} & \mathbf{-2,8062} \\ \mathbf{0} & \mathbf{0} & \mathbf{2,3717} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{4,6771} \end{bmatrix}$$

A matriz **M** tem **t=4** linhas, uma para cada idade mensurada, e **k=4** colunas, sendo uma para a média, ajuste linear, quadrático e cúbico, respectivamente, e é dada por

$$\mathbf{M} = \begin{bmatrix} \mathbf{1} & \mathbf{-1} & \mathbf{1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-0,3333} & \mathbf{0,1111} & \mathbf{-0,0370} \\ \mathbf{1} & \mathbf{0,3333} & \mathbf{0,1111} & \mathbf{0,0370} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}.$$

A matriz resultante f é definida por **M**L, resultando em

$$f = \begin{bmatrix} \mathbf{0,7071} & \mathbf{-1,2247} & \mathbf{1,5811} & \mathbf{-1,8709} \\ \mathbf{0,7071} & \mathbf{-0,4082} & \mathbf{-0,5271} & \mathbf{0,7622} \\ \mathbf{0,7071} & \mathbf{0,4082} & \mathbf{0,5271} & \mathbf{-0,7622} \\ \mathbf{0,7071} & \mathbf{1,2247} & \mathbf{1,5811} & \mathbf{1,8709} \end{bmatrix}$$

e a matriz coeficiente **K** é estimada por $\hat{\mathbf{K}} = f^{-1}\hat{\mathbf{G}}(f^{-1})'$, obtendo-se

$$\hat{\mathbf{K}} = \begin{bmatrix} \mathbf{4,3601} & \mathbf{0,5287} & \mathbf{-0,2163} & \mathbf{-0,0403} \\ \mathbf{0,5287} & \mathbf{0,4747} & \mathbf{0,0045} & \mathbf{-0,1074} \\ \mathbf{-0,2163} & \mathbf{0,0045} & \mathbf{0,0675} & \mathbf{0,0040} \\ \mathbf{-0,0403} & \mathbf{-0,1074} & \mathbf{0,0040} & \mathbf{0,0760} \end{bmatrix}$$

A matriz de coeficientes da função de covariância é dada por:

$$\mathbf{C} = L\hat{\mathbf{K}}L' = \begin{bmatrix} \mathbf{2,464} & \mathbf{0,5424} & \mathbf{-0,4894} & \mathbf{-0,1483} \\ \mathbf{0,5424} & \mathbf{2,0473} & \mathbf{-0,0134} & \mathbf{-1,6101} \\ \mathbf{-0,4894} & \mathbf{-0,0134} & \mathbf{0,3797} & \mathbf{0,0443} \\ \mathbf{-0,1483} & \mathbf{-1,6101} & \mathbf{0,0443} & \mathbf{1,6579} \end{bmatrix}.$$

Assim, a função de covariância é

$$\mathbf{f(x_l, x_m) = 2,464 + 0,542(x_l + x_m) + 2,047x_l x_m - 0,489(x_l^2 + x_m^2) - 0,013(x_l^2 x_m + x_l x_m^2) + 0,379x_l^2 x_m^2}$$
$$\mathbf{- 0,148(x_l^3 + x_m^3) - 1,610(x_l^3 x_m + x_l x_m^3) + 0,044(x_l^3 x_m^2 + x_l^2 x_m^3) + 1,657x_l^3 x_m^3.}$$

Usando esta função podemos estimar a covariância entre as combinações de idades representada em $\hat{\mathbf{G}}$, e pela interpolação podemos também estimar a covariância entre quaisquer duas idades que nunca foram mensuradas. Por exemplo, a covariância entre o peso a 3,5 ($\mathbf{x_1}$=0) e 4,0 ($\mathbf{x_2}$=0,3333) semanas de idade pode ser estimada por:

$$\mathbf{f(0, 0,3333) = 2,464 + 0,542(0,3333) - 0,489(0,3333^2) - 0,148(0,3333^3) = 2,58.}$$

No exemplo acima, considerou-se o ajuste completo, ou seja, a ordem do polinômio é igual ao número de idades avaliadas. Para se obter um ajuste de polinômios

de menor ordem (**k < t**), KIRKPATRICK et al. (1994) descreveram um procedimento de quadrados mínimos generalizados para determinar os coeficientes da função de covariância, a partir de uma matriz de covariâncias pré-estimada.

No método de ajuste reduzido, se produz um ajuste da função de covariância baseado em **k** funções ortogonais, em que **k** é menor que **t**, e **t** é a dimensão de $\hat{\mathbf{G}}$, isto é, o número de idades mensuradas. O método consiste em dois passos: no primeiro , a função de covariância candidata é construída por meio de quadrados mínimos generalizados para ajustar uma função ortogonal mais simples e no segundo passo, a função candidata é testada para a verificação de sua consistência estatística com a matriz $\hat{\mathbf{G}}$. O ajuste é testado usando-se a distribuição de c^2. Se a função diferir estatisticamente de $\hat{\mathbf{G}}$, considera-se uma estimativa de função de covariância reduzida mais complexa em relação à anterior. A nova função estimada é então testada. Se o teste for não significativo, a função é considerada consistente com $\hat{\mathbf{G}}$, caso contrário, o processo se repete iterativamente.

Estimação de Funções de Covariância por meio do Método da Máxima Verossimilhança Restrita

Como discutido anteriormente, o modelo de função de covariância requer que as estimativas de matrizes de covariância entre as medidas nas **t** idades observadas estejam disponíveis. Na prática, é preferível fazer o ajuste reduzido diretamente dos dados observados. Além disso, é desejável fazer de modo seqüencial, ou seja, testando algumas ordens de polinômios até que a estrutura da matriz de variâncias e covariâncias possa ser modelada adequadamente com um menor número de parâmetros. Isto pode ser feito prontamente por meio do método de estimação da máxima verossimilhança restrita (REML).

Um método para estimar a função de covariância por meio do método REML foi proposto por MEYER e HILL (1997). A maior vantagem deste método está no fato de que o método de estimação REML permite a garantia do espaço paramétrico. Sendo assim, a matriz estimada de coeficientes **K** é positiva definida, o que não é o caso quando se usa o método de quadrados mínimos generalizados de KIRKPATRICK et al. (1990).
A seguir será feita uma abordagem sobre o método de MEYER e HILL (1997).

Seja $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$ o modelo linear geral multi-característica, com **y**, **b**, **u** e **e** sendo os vetores de observações, efeitos fixos, efeitos aleatórios e erros residuais, respectivamente, e **X** e **Z** as matrizes de incidência relacionadas a **b** e **u**, respectivamente. Para este modelo, **u** inclui o vetor de efeitos genéticos aditivos do indivíduo (**a**), **e** pode conter efeitos aleatórios adicionais, como efeitos de ambiente permanente.

Seja $\mathbf{V(u)=G}$, $\mathbf{V(e)=R}$ e $\mathbf{Cov(u,e)=0}$, tal que $\mathbf{V(y)=V=ZGZ'+R}$, assumindo uma distribuição normal multivariada, ou seja, $\mathbf{y \sim N(Xb,\ V)}$, Sendo **t** diferentes idades mensuradas por indivíduo, com uma única mensuração por idade, considerando, para simplificar, um modelo básico com efeitos genéticos aditivos sendo os únicos efeitos aleatórios ajustados (**u=a**) e assumindo que todos os indivíduos tenham observações para todas as idades. Então **u** é um vetor de efeitos genéticos aditivos com Var(**u**) = **G**, com $\mathbf{G=A \ddot{A} G_0}$, em que **A** é a matriz de parentesco entre os indivíduos e $\mathbf{G_0}$ é a matriz de ordem **t** com variâncias e covariâncias genéticas entre as **t** características. Se não há valores perdidos, as matrizes de incidência são iguais para todas as características. Ainda tem-se que, Var(**e**) = $\mathbf{R = I \ddot{A} R_0}$, em que **I** é uma matriz identidade de ordem igual ao número de observações e R_o representa a variância residual. O logaritmo da função de máxima verossimilhança restrita é dado por

$$\mathbf{lnL = -1/2 \left[const + N \ln|R_0| + N_a \ln|G_0| + t \ln |A| + \ln |C| + y'Py \right]}$$

em que $\mathbf{N_a}$ é o número de indivíduos na análise, **C** é uma matriz de coeficientes para as equações de modelo misto e **y'Py** são as somas de quadrado residuais. MEYER e HILL

(1997) sugeriram que para a estimação de uma função de covariância, o logaritmo da função de máxima verossimilhança restrita poderia ser reescrito como:

$$\mathbf{lnL = -1/2\ [\ const + N\ ln|K_e| + N_a\ ln|K_a| + (N+N_a)\ ln|\ ff¢| + t\ ln\ |A| + ln\ |C| + y'Py]}$$

em que $fK_a f¢$ e $fK_e f¢$ são funções de covariância para $\mathbf{G_0}$ e $\mathbf{R_0}$, respectivamente. Ou seja, no lugar de $\mathbf{G_0}$ e $\mathbf{R_0}$, foram inseridas as matrizes de covarâncias entre os coeficientes da função de covariância, $\mathbf{K_a}$ e $\mathbf{K_e}$. Para o cálculo de **y'Py** e **ln|C|** os autores utilizaram as equações de modelos mistos para um modelo multi-característica. Uma alteração no logaritmo da função foi sugerida pelos autores, e é dada por:

$$\mathbf{lnL = -1/2[const + N\,ln|\,fK_r f¢ + Diag\{s_e^2\}|\, + N_s\ ln|\,K_a\,| + (N_a)\,ln|\,ff¢|\, + t\,ln|\,A\,| + ln|\,C\,| + y¢Py\ ,}$$

onde $\mathbf{K_r}$ representa a matriz de coeficientes de função de covariância para o efeito de ambiente permanente e **Diag{** s_e^2 **}** a matriz de variâncias de ambiente temporário. Deste modo, foi possível separar a variância do efeito ambiental em variância de efeito de ambiente permanente e variância de efeito de ambiente temporário.

A vantagem deste procedimento sobre o modelo multi-característica, com **t** características, é que a dimensão do espaço paramétrico é reduzida, e há somente **k(k+1)/2** parâmetros para serem estimados, ao invés de **t(t+1)/2**, onde **k** é a ordem do polinômio empregado.

Em modelos de avaliação genética com base em medidas repetidas, comumente há, no mínimo, três componentes aleatórios: efeito de indivíduo, ambiente permanente e ambiente temporário. Em cada um destes componentes existe uma diferente estrutura de covariância. O modelo estatístico pode, então, ser definido com os efeitos aleatórios genético aditivo e de ambiente permanente representados por funções de covariância.

Considere o modelo:

$$\mathbf{y_i} = m + \mathbf{u_i} + \mathbf{p_i} + e_i$$

em que $\mathbf{u_i}$ é um vetor com efeitos genéticos aditivos para a observação mensurada no indivíduo **i**, e $\mathbf{p_i}$ e e_i são vetores com efeitos permanente e temporário de ambiente, respectivamente. Então:

$$\mathbf{Var(u_i) = G_0; Var\ (p_i) = P_0}\ \text{e}\ \mathbf{Var}(e_i) = \mathbf{I}\,\sigma^2_e\,\mathbf{.}$$

Se todos os indivíduos possuem mensurações nas mesmas idades, $\mathbf{G_0}$ e $\mathbf{P_0}$ possuem as mesmas dimensões. Em um modelo multi-característica, a matriz de covariância residual torna-se:

$$\mathbf{Var(e) = Var\ (p_i} + e_i\mathbf{) = P_0 + I}\,\sigma^2_e\,\mathbf{= R_0}.$$

Assim, admite-se que as medidas de erros são independentes entre as idades, e uma função de covariância é descrita somente para o efeito genético aditivo e de ambiente permanente. Assumindo-se uma função de covariância ajustada pelos polinômios de Legendre, por exemplo, para o efeito genético aditivo e de ambiente permanente e com a mesma ordem de ajuste (VAN DER WERF e SCHAEFFER, 1997), as matrizes de covariância genética aditiva e de ambiente permanente podem ser representadas, por funções de covariância fornecidas, respectivamente, por:

$$\mathbf{G_0} = \phi\mathbf{K_a}\phi'$$
$$\mathbf{P_0} = \phi\mathbf{K_p}\phi'\mathbf{.}$$

Diante disso, o modelo misto pode ser reescrito substituindo $\mathbf{u_i}$ por $\phi_i\,\mathbf{a_i}$ e $\mathbf{p_i}$ por $\phi_i\,\mathbf{p_i}$, como segue:

$$\mathbf{y_i} = \mu + \phi_i\,\mathbf{a_i} + \phi_i\,\mathbf{p_i} + e_i\,\mathbf{.}$$

Se a função de covariância é estimada pelo método de ajuste completo (o grau do polinômio é igual ao número de mensurações), os modelos anteriormente descritos são equivalentes, possuindo a mesma esperança e variância, ou seja,

$$\mathbf{Var}(\phi_i \mathbf{a}_i) = \phi_i \mathbf{Var}(\mathbf{a}_i) \phi_i' = \phi_i \mathbf{K}_a \phi_i' = \mathbf{G}_0 = \mathbf{Var}(\mathbf{u}_i);$$
$$\mathbf{Var}(\phi_i \mathbf{p}_i) = \phi_i \mathbf{Var}(\mathbf{p}_i) \phi_i' = \phi_i \mathbf{K}_p \phi_i' = \mathbf{P}_0 = \mathbf{Var}(\mathbf{p}_i).$$

Se os indivíduos têm registros em idades diferentes, o modelo de função de covariância requer uma matriz ϕ_i diferente para cada conjunto de idades diferentes mensuradas, mas o coeficiente de regressão tem a mesma estrutura de covariância ($\mathbf{K}_a$ e $\mathbf{K}_p$) para cada indivíduo, independente do conjunto de idades mensuradas. O número de efeitos genéticos aditivos preditos para cada indivíduo no modelo multi-característica é igual ao número de idades em que o indivíduo foi mensurado. No modelo de análise que emprega a função de máxima verossimilhança restrita reparametrizada (MEYER e HILL, 1997) o número de soluções que levam à predição do efeito genético aditivo por indivíduo é igual à ordem do ajuste da função de covariância, ou seja, à ordem de $\mathbf{K}_a$. Nesta situação, tem-se um modelo de regressão em que os dados são regredidos nos polinômios de Legendre com a variável regressora em ϕ e os coeficientes de regressão em **a** e **p**. Os coeficientes de regressão não são os mesmos para cada indivíduo, e são chamados de coeficientes de regressão aleatória, com $\mathbf{Var(a)} = \mathbf{K}_a$ e $\mathbf{Var(p)} = \mathbf{K}_p$. Reescrevendo o modelo misto multi-característica para um modelo misto com função de covariância, tem-se o formato de um modelo unicaracterístico de regressão aleatória, com cada efeito aleatório tendo **k** coeficientes de regressão aleatória. Um modelo para **n** observações em **q** indivíduos pode ser reescrito como:

$$\mathbf{y} = \mathbf{Xb} + \sum_{j=0}^{k-1} \mathbf{Z}_j \mathbf{a}_j + \sum_{j=0}^{k-1} \mathbf{Z}_j \mathbf{p}_j + \mathbf{e}$$

em que $\mathbf{Z}_j$ são matrizes **n** x **q** para o **i**-ésimo polinômio, $\mathbf{a}_j$ e $\mathbf{p}_j$ são vetores com coeficientes de regressão aleatória para todos os indivíduos, referente aos efeitos genético aditivo e permanente de ambiente, respectivamente. A matriz $\mathbf{Z}_j$ contém a variável

regressora, ou seja, os polinômios em f. Como cada indivíduo tem **k** coeficientes em **a** e **k** coeficientes em **p** (admitindo o mesmo ajuste para **a** e **p**), a matriz **Z** pode ser escrita como uma matriz bloco diagonal de ordem **n** x **kq**, com um bloco $\mathbf{Z}_i = f_i = \mathbf{M}_i L$ para cada indivíduo **i**. O modelo misto pode ser expresso como:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}^*\mathbf{a} + \mathbf{W}^*\mathbf{p} + e$$

com $\mathbf{a}' = [\mathbf{a}_1', ..., \mathbf{a}_q']$ e $\mathbf{p}' = [\mathbf{p}_1', ..., \mathbf{p}_q']$, com $\mathbf{a_i}$ e $\mathbf{p_i}$ sendo o conjunto de coeficientes de regressão aleatória para o indivíduo **i**. Se todos os indivíduos têm mensurações nas mesmas idades, todas $\mathbf{Z_i}$ são iguais e $\mathbf{Z}^* = \mathbf{I_q} \otimes f$. A variância e covariância dos efeitos aleatórios podem ser escritas como: $\mathbf{Var(a)} = \mathbf{A} \otimes \mathbf{K_a}$ e $\mathbf{Var(p)} = \mathbf{I} \otimes \mathbf{K_p}$ e **Cov(a,p) =** Æ, em que $\mathbf{K_a}$ e $\mathbf{K_p}$ são os coeficientes para a função de covariância para os efeitos genético aditivo e permanente de ambiente.

As equações de modelo misto da regressão aleatória com função de covariância, têm uma estrutura similar com o modelo de repetibilidade, exceto que mais coeficientes são gerados por meio das variáveis regressoras polinômicas de f, que são incorporados em **Z**. Na parte das equações de modelo misto para o efeito genético aditivo, há, para cada indivíduo, uma diagonal em blocos $f_i' f_i + \mathbf{a}^{ii} s_e^2 \mathbf{K}_a^{-1}$ e fora da diagonal blocos $\mathbf{a}^{ij} s_e^2 \mathbf{K}_a^{-1}$, sendo $\mathbf{a}^{ij}$ o **i**,**j**-ésimo elemento da inversa da matriz de coeficientes de parentesco dos indivíduos ($\mathbf{A}^{-1}$). A parte do efeito aleatório de ambiente permanente é uma diagonal em blocos com blocos diagonais iguais a $f'f_i + s_e^2 \mathbf{K}_p^{-1}$. As equações de modelo misto são então representadas por:

$$\begin{bmatrix} \mathbf{X}_i'\mathbf{X}_i & \cdots & \mathbf{X}_i' f_i & \cdots & \mathbf{X}_i' f_i & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_i'\mathbf{X}_i & \cdots & f_i' f_i + \mathbf{a}^{ii} s_e^2 \mathbf{K}_a^{-1} & \cdots & f_i' f_i & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_i'\mathbf{X}_i & \cdots & f_i' f_i & \cdots & f_i' f_i + s_e^2 \mathbf{K}_p^{-1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \vdots \\ \mathbf{a}_i \\ \vdots \\ \mathbf{p}_i \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i'\mathbf{y}_i \\ \vdots \\ f_i'\mathbf{y}_i \\ \vdots \\ f_i'\mathbf{y}_i \\ \vdots \end{bmatrix},$$

em que o subscrito i se refere à parte das equações referente ao indivíduo i (VAN DER WERF e SCHAEFFER, 1997).

Os coeficientes polinomiais geram, portanto, nas equações de modelo misto, uma matriz de coeficientes mais densa, devido à presença, agora, das variáveis regressoras. Em um modelo misto multi-característica estimam-se **t** valores genéticos por indivíduo, ou seja, um para cada característica. No modelo de regressão aleatória usando uma função de covariância, têm-se 2**k** coeficientes de regressão a serem estimados por indivíduo, sendo **k** coeficientes para o efeito genético aditivo e **k** coeficientes para o efeito de ambiente permanente. A similaridade entre um modelo de regressão aleatória, usando função de covariância com o modelo misto multi-característica, pode ser demonstrada.

Sendo o modelo misto multi-característica $\mathbf{y=Xb + Zu +e}$, com $\mathbf{Var(y) = ZGZ'+R}$, **u** é um vetor de efeito genético aditivo do indivíduo, com $\mathbf{Var(u) = G}$. Se **u**, assim como **y**, são ordenados por indivíduo, bem como ordenam-se as características dentro de indivíduo, **Z** é uma matriz diagonal em blocos, com blocos diagonais $\mathbf{Z_j}$ pertencente a cada indivíduo, e $\mathbf{Z_j}$ é uma matriz $\mathbf{n_j \times t}$, sendo $\mathbf{n_j}$ o número de características mensuradas para o indivíduo **j**. Definindo $\mathbf{Var(u) = G = A \otimes G_0}$ e $\mathbf{Var(e) = R = I \otimes Z_i R_0 Z_i'}$, com $\mathbf{G_0}$ e $\mathbf{R_0}$ de ordem **t**, se não há valores perdidos, as matrizes de incidência são iguais para cada uma das características, $\mathbf{Z_i = I_t}$ e $\mathbf{R = I \otimes R_0}$.

Supondo um modelo de regressão aleatória usando função de covariância com o grau do polinômio **k** (**k = t**), com todos indivíduos possuindo mensurações nas **t** idades, é possível comparar este modelo ao modelo misto multi-característica com todos indivíduos apresentando mensurações nas **t** idades. Se a função de covariância é definida de forma que $\mathbf{K_a}$ e $\mathbf{K_p}$ são estimados a partir de $\mathbf{G_0}$ e $\mathbf{R_0}$-**Var**(e), respectivamente, em que **Var**(e)= $\mathbf{I}\sigma_e^2$ é a variância do efeito temporário de ambiente, então:

$$\mathbf{Var(Z^{*}a)} = \mathbf{A} \otimes (\mathbf{K_a} \otimes \mathbf{I_q}) = \mathbf{A} \otimes \mathbf{G_0} = \mathbf{Var(Zu)} \quad \mathbf{e}$$
$$\mathbf{Var(Z^{*}p)} = \mathbf{I_q} \otimes (\mathbf{K_p} \otimes \mathbf{I_q}) + \mathbf{I}\sigma_e^2 \otimes \mathbf{I_q} \otimes \mathbf{R_0} = \mathbf{Var(e)}.$$

O vetor de valores genéticos aditivos de características múltiplas para o indivíduo **i**, $\mathbf{u_i}$, pode ser, portanto, denotado por $f\,\mathbf{a}_i$, ou seja, os coeficientes de regressão aleatória são pré-multiplicados por f. Com a ordem do ajuste dos polinômios **k** = **t**, o modelo misto de características múltiplas é exatamente igual ao modelo de regressão aleatória com função de covariância. Se **k** < **t** (por exemplo, quando as mensurações nas **t** idades são altamente correlacionadas), a estrutura de covariância descrita pela função de covariância será mais suave e provavelmente mais correta que a estrutura de covariância de um modelo multi-característica com **t** características, com **t(t+1)/2** estimadores. Um modelo multi-característica para muitas variáveis é numericamente menos estável se a matriz de covariância tem muitos autovalores próximos de zero e a inversa de tais matrizes pode ser imprecisa.

Em um modelo de função de covariância, a informação para cada indivíduo em cada efeito aleatório é representada por **k** coeficientes de regressão, e as soluções para os valores genéticos em cada conjunto de idades consideradas podem ser geradas por $\mathbf{u}_i = f\,\mathbf{a}_i = \mathbf{M}_i\mathbf{a}_i$. O modelo de regressão aleatória com função de covariância é mais flexível que um modelo multi-característica, sendo possível lidar com mensurações em algum estágio definido e as soluções são aproximadamente iguais às soluções do modelo de características múltiplas com **t** características. A aproximação depende da acurácia da **k**-ésima ordem do ajuste da função de covariância na **t**-dimensional matriz de (co)variância.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, L.G. 2003. Modelos de dimensão infinita aplicados a características de crescimento de bovinos da raça Nelore. Jaboticabal: UNESP, 2003. 83p. Tese (LIVRE-DOCENTE)- Faculdade de Ciências Agrárias e Veterinárias. 2003.

ARAÚJO, C. V. 2003. Modelos de regressão aleatória para avaliação genética da produção de leite na raça Holandesa. Viçosa; UFV, 2003. 85p. Tese (Doutorado)- Universidade Federal de Viçosa. 2003.

BROTHESTONE, S; WHITE, I. M. S.; MEYER, K. 2000. Genetic modeling of daily milk yield using orthogonal polynomials and parametric curves. *Animal Science*, vol 70, p.407-415, 2000.

COBUCI, J. A. 2002. Uso de modelos de regressão aleatória na avaliação da persistência na lactação de animais da raça Holandesa**.** Viçosa: UFV, 2002, 99 p. Tese (Doutorado em Zootecnia) – Universidade Federal de Viçosa, MG, 2002.

COSTA, C. N.; MELO, C. M. R.; MACHADO, C. H. C. et al. 2002. Avaliação de funções polinomiais para ajuste da produção no dia de controle de primeiras lactações de vacas Gir com modelos de regressão aleatória. In: Reunião Anual da Sociedade Brasileira de Zootecnia, 39., 2002, Recife. *Anais* ... Recife: SBZ, 2002, (CD-ROM)

EL FARO, L. 2002 Estimação de componentes de (co)variância para a produção de leite no dia de controle de primeiras lactações de vacas Caracu, aplicando-se “test day models” de dimansão finita e modelos de regressão aleatória**.** Jaboticabal: UNESP, 2002, 102 p. Tese (Doutorado em Zootecnia) – Universidade Estadual Paulista, SP, 2002.

HENDERSON JUNIOR, C. R. 1982. Analysis of covariance in the mixed model: higher level, nonhomogeneus, and random regressions. *Biometrics* v.38, p.623-640, 1982.

JAMROZIK, J., SCHAEFFER. L. R. 1997. Estimates of genetic parameters for a test day model with random regression for yield traits of first lactation Holstein**.** *Journal of Dairy Science*, vol. 80, n. 4, p. 762-770, 1997.

JAMROZIK, J., KISTEMAKER, G.J., DEKKERS, J.C.M., SCHAEFFER, L.R. 1997. Comparison of possible covariates for use in a random regression model for analyses of test day yields. *Journal of Dairy Science*, vol. 80, n.8, p.2550-2556, 1997.

KIRKPATRICK, M. HECKMAN, N. 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *Journal Mathematic Biological.*, vol. 27, p. 429-450, 1989.

KIRKPATRICK, M.; LOFSVOLD, D.; BULMER, M. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, vol. 24, n. 3, p. 979-993, 1990.

KIRKPATRICK, M.; HILL, W. G.; THOMPSON, R. 1994. Estimating the covariance structure of traits durink growth and aging, illustrated with lactations in dairy cattle. *Genetic Research*, v.64, p.57-69, 1994.

MEYER, K.; HILL, W. G. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or "repetead" records by restricted maximum likelihood. *Livestock Production Science*, vol. 47, n. 3, p. 185-200, 1997.

MEYER, K. 2004. Scope for a random regression model in genetic evaluation of beef cattle for growth. *Livestock Production Science,* vol. 86, p. 69-83, 2004.

OLORI, V.E., HILL, W.G., MCGUIRK, B.J. et al. 1999. Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livestock Production Science,* vol. 61, n. 53, p. 63, 1999.

REKAYA, R., CARABANO, M.J., TORO, M.A. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. *Livestock Production Science,* vol. 57, n. 3, p.203-217, 1999.

RESENDE, M. D. V.; REZENDE, G. D. S. P.; FERNANDES, J. S. C. 2001. Regressão aleatória e funções de covariância na análises de medidas repetidas. *Revista de Matemática e Estatística*, vol.19:21-40, 2001.

RESENDE, M. D. V. 2002. Genética biométrica e estatística no melhoramento de plantas perenes, 1 ed, Brasília: EMBRAPA Informação Tecnológica, 2002, 975p.

SAKAGUTI, E.S. 2000. Funções de covariância e modelos de regressão aleatória na avaliação genética do crescimento de bovinos jovens da raça Tabapuã. Viçosa; UFV, 2000. 81p. Tese (Doutorado)- Universidade Federal de Viçosa. 2000.

SCHAEFFER L. R. 2000. Random regression models. http://nitro.biosci.arizona.edu/zbook/book.htlm. Acessado em dezembro de 2001.

SCHAEFFER, L. R.; DEKKERS, J. C. M. 1994. Random regression in animal models for test day production in dairy cattle. In: World congress genetic applied livestock production, 5., 1994, Guelph, ON, Canada, *Procedings* … Guelph, 1994. p.443-446.

SCHAEFFER, L. R. 2004. Aplication of random regression models in animal breeding *Livestock Production Science*, vol. 86, p. 35 - 45, 2004.

TIJANI, A., WIGGANS, G.R., VAN TASSELL, C.P. 1998. Use of (co)variance functions to describe (co)variance for test day yield. *Journal of Dairy Science*, vol. 82, n. 1, p. 226, 1998.

VAN DER WERF, J. H. J.; GODDARD, M. E.; MEYER, K. 1998. The use of covariance functions and random regression for genetic evaluation of milk production basead on test day records. *Journal of Dairy Science*, vol 81, n 12, p.3300-3308, 1998.

VAN DER WERF, J. H. J.; SCHAEFFER, L. 1997. Random regression in animal breeding. Course Notes, Ontario: University of Guelph, 1997, p. 70.

Capítulo 1

Aplicação de um modelo de regressão aleatória

Modelos de regressão aleatória são utilizados para a avaliação de dados com estrutura do tipo longitudinal, isto é, quando são feitas mensurações repetidas ao longo do tempo, no mesmo indivíduo. Suponha a seguinte estrutura de dados hipotéticos apresentados na Tabela 1, exemplificando mensurações da variável diâmetro à altura do peito (DAP) em eucalipto.

Tabela 1: Dados referentes à variável diâmetro à altura do peito (DAP) para três indivíduos mensurados em diferentes idades.

Indivíduo	Pai	Mãe	Idade (meses)	Local	DAP(cm)
1	4	6	12	1	15,1
1	4	6	36	1	25,4
1	4	6	48	1	42,1
2	4	7	11	2	13,8
2	4	7	52	2	45,4
3	5	7	14	2	16,2
3	5	7	35	2	23,9
3	5	7	47	2	41,8

Neste conjunto de dados, foram avaliados três indivíduos. Os indivíduos 1 e 3 possuem informação de avaliação em três idades (indivíduo 1 aos 12, 36 e 48 meses e o indivíduo 3 aos 14, 35 e 47 meses), ao passo que o indivíduo 2 possui informações apenas aos 11 e 52 meses. Uma das vantagens da utilização de modelos de regressão aleatória, é que os indivíduos não precisam ser necessariamente todos mensurados nas mesmas idades. Note que neste exemplo, as idades em que os

indivíduos foram avaliados foram todas diferentes. Observe também que a menor idade de mensuração foi aos 11 meses (indivíduo 2) e a maior idade foi aos 52 meses (indivíduo 2). Portanto, o estudo desses dados permitirá que o pesquisador faça inferências a respeito de qualquer um dos três indivíduos em qualquer ponto no intervalo compreendido entre 11 e 52 meses de idade.

Quando se adota um modelo linear como o multi-característica ou repetibilidade, por exemplo, as equações de modelos mistos podem ser utilizadas para a obtenção das soluções que permitem fazer a classificação dos indivíduos para ordenamento e seleção. A estrutura de variâncias e covariâncias requeridas pelas equações de modelos mistos é fornecida por uma matriz de variâncias e covariâncias entre as idades. Isto porque nestes modelos trabalha-se com pontos arbitrários dentro de um intervalo estabelecido, ou seja, os dados amostrais são dispostos em uma escala descontínua.

Em modelos de regressão aleatória, trabalha-se admitindo que os dados amostrais estão dispostos em uma escala contínua, apresentado-se no nível de uma trajetória no tempo, portanto, uma função que descreve a estrutura de variâncias e covariâncias entre idades deve ser adotada. Tal função é denominada função de covariância, e permite descrever todas as variâncias e covariâncias para qualquer ponto em uma trajetória. Várias funções de covariância podem ser utilizadas. Atualmente, os polinômios de Legendre têm sido bastante adotados porque são fáceis de calcular e fáceis de interpretar, pois são definidos no intervalo de –1 a 1. Além disso, não requerem um conhecimento prévio sobre a forma da curva que descreve a trajetória.

A utilização de funções de covariância utilizando os polinômios de Legendre foi descrita por KIRKPATRICK et al. (1990), e é feita com base nos seguintes passos:

1. Todas as idades que são feitas as avaliações devem ser padronizadas no intervalo –1 a 1.

2. Será obtida uma matriz de polinômios de Legendre, que será determinada de acordo com o grau do polinômio adotado, considerando-se o número de idades avaliadas. Quando se faz o ajuste completo, o grau do polinômio é igual ao número de idades avaliadas, e ao contrário, ao se fazer um ajuste reduzido, o grau do polinômio utilizado é menor que o número de idades avaliadas.
3. Uma matriz que associa os polinômios de Legendre às idades padronizadas é obtida.
4. Uma matriz com os coeficientes da função de covariância pode ser estimada, e permite estimar todas as variâncias e covariâncias entre idades, no intervalo compreendido entre a menor e a maior idade avaliada.

Utilizando os dados fornecidos no exemplo da Tabela 1, podemos padronizar as idades (Passo 1) por meio da seguinte expressão:

$$\mathbf{a}_i^* = \mathbf{u} + \frac{\mathbf{v} - \mathbf{u}}{\mathbf{a}_{max} - \mathbf{a}_{min}}\left(\mathbf{a}_i - \mathbf{a}_{min}\right).$$

$\mathbf{a}_{max}$ e $\mathbf{a}_{min}$ são a maior e menor idade mensurada, respectivamente. Para o exemplo em questão, $\mathbf{a}_{max} = \mathbf{52}$ e $\mathbf{a}_{min} = 11$. **u** e **v** representam o intervalo para o qual as idades serão padronizadas, sendo representados por –1 e 1, respectivamente. A título de ilustração, a padronização da idade 12 meses é obtida por:

$$\mathbf{a}_{12} = \mathbf{-1} + \frac{\mathbf{1 - (-1)}}{\mathbf{52 - 11}}\left(\mathbf{12 - 11}\right) = \mathbf{-0{,}9512}.$$

Logicamente, a menor idade fica padronizada em –1, e a maior idade fica padronizada em 1. As idades padronizadas para o exemplo em questão, são:

Idade de avaliação	Idade padronizada
12	-0,9512
36	0,2195
48	0,8048
11	-1,0000
52	1,0000
14	-0,8536
35	0,1707
47	0,7561

A matriz de idades padronizadas poderá então ser definida de acordo com o grau do polinômio que será ajustado. Considerando que sempre se busca o polinômio de menor ordem possível (não faz sentido ajustar um polinômio completo, dado que normalmente muitas idades são avaliadas), vamos considerar para este exemplo, um ajuste quadrático, dado que o número máximo de idades avaliadas foi de três (nos indivíduos 1 e 3). Então, a matriz de idades padronizadas irá incluir na primeira coluna o intercepto, na segunda coluna as idades padronizadas e na terceira coluna o quadrado das idades padronizadas, como segue:

$$\mathbf{M} = \begin{bmatrix} \mathbf{1} & \mathbf{-0,9512} & \mathbf{0,9048} \\ \mathbf{1} & \mathbf{0,2195} & \mathbf{0,0481} \\ \mathbf{1} & \mathbf{0,8048} & \mathbf{0,6478} \\ \mathbf{1} & \mathbf{-1,0000} & \mathbf{1,0000} \\ \mathbf{1} & \mathbf{1,0000} & \mathbf{1,0000} \\ \mathbf{1} & \mathbf{-0,8536} & \mathbf{0,7287} \\ \mathbf{1} & \mathbf{0,1707} & \mathbf{0,0291} \\ \mathbf{1} & \mathbf{0,7561} & \mathbf{0,5716} \end{bmatrix}.$$

Os polinômios de Legendre (Passo 2) podem ser obtidos por meio da seguinte expressão:

$$\mathbf{P_{n+1}(x) = \frac{1}{n+1}\left[(2n+1)x\,P_n(x) - n\,P_{n-1}(x)\right].}$$

Essa quantidade pode ser normalizada por:

$$\mathbf{f_n(x) = \sqrt{\frac{2n+1}{2}}P_n(x).}$$

Inicialmente, definimos $\mathbf{P_0(x) = 1}$ e $\mathbf{P_1(x) = x}$. Assim, o cálculo do polinômio de ordem 1 é obtido por:

$$\mathbf{f_0(x) = \sqrt{\frac{1}{2}}P_0(x) = \sqrt{\frac{1}{2}}(1) = 0{,}7071.}$$

O polinômio de ordem 2 pode ser obtido por:

$$\mathbf{f_1(x) = \sqrt{\frac{3}{2}}P_1(x) = \sqrt{\frac{3}{2}}x = 1{,}2247x.}$$

O polinômio de ordem 3 é obtido por:

$$\mathbf{P_{1+1}(x) = P_2(x) = \frac{1}{2}\left[3x\,P_1(x) - P_0(x)\right]}$$

$$\mathbf{P_2(x) = \frac{1}{2}(3x.x - 1)}$$

$$\mathbf{P_2(x) = \frac{1}{2}(3x^2 - 1)}$$

$$\mathbf{P_2(x) = \frac{3x^2}{2} - \frac{1}{2},}$$

então,

$$f_2(\mathbf{x}) = \sqrt{\frac{5}{2}}\left(\frac{3\mathbf{x}^2}{2} - \frac{1}{2}\right) = \mathbf{-0{,}7906 + 2{,}3717x^2}.$$

Obtidos os polinômios, a matriz de polinômios, considerando o ajuste de ordem três (quadrático), fica assim definida:

$$L = \begin{bmatrix} \mathbf{0{,}7071} & \mathbf{0} & \mathbf{-0{,}7906} \\ \mathbf{0} & \mathbf{1{,}2247} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2{,}3717} \end{bmatrix},$$

Agora podemos fazer a associação da matriz de polinômios L, à matriz de idades padronizadas **M** (Passo 3).

$$f = \mathbf{M}L = \begin{bmatrix} \mathbf{1} & \mathbf{-0{,}9512} & \mathbf{0{,}9048} \\ \mathbf{1} & \mathbf{0{,}2195} & \mathbf{0{,}0481} \\ \mathbf{1} & \mathbf{0{,}8048} & \mathbf{0{,}6478} \\ \mathbf{1} & \mathbf{-1{,}0000} & \mathbf{1{,}0000} \\ \mathbf{1} & \mathbf{1{,}0000} & \mathbf{1{,}0000} \\ \mathbf{1} & \mathbf{-0{,}8536} & \mathbf{0{,}7287} \\ \mathbf{1} & \mathbf{0{,}1707} & \mathbf{0{,}0291} \\ \mathbf{1} & \mathbf{0{,}7561} & \mathbf{0{,}5716} \end{bmatrix} \begin{bmatrix} \mathbf{0{,}7071} & \mathbf{0} & \mathbf{-0{,}7906} \\ \mathbf{0} & \mathbf{1{,}2247} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2{,}3717} \end{bmatrix}$$

$$f = \begin{bmatrix} \mathbf{0{,}7071} & \mathbf{-1{,}1649} & \mathbf{1{,}3553} \\ \mathbf{0{,}7071} & \mathbf{0{,}2688} & \mathbf{-0{,}6763} \\ \mathbf{0{,}7071} & \mathbf{0{,}9857} & \mathbf{0{,}7458} \\ \mathbf{0{,}7071} & \mathbf{-1{,}2247} & \mathbf{1{,}5811} \\ \mathbf{0{,}7071} & \mathbf{1{,}2247} & \mathbf{1{,}5811} \\ \mathbf{0{,}7071} & \mathbf{-1{,}0454} & \mathbf{0{,}9377} \\ \mathbf{0{,}7071} & \mathbf{0{,}2090} & \mathbf{-0{,}7214} \\ \mathbf{0{,}7071} & \mathbf{0{,}9259} & \mathbf{0{,}5652} \end{bmatrix}.$$

Note que as três primeiras linhas da matriz F referem-se ao indivíduo 1, a quarta e quinta linha se referem ao indivíduo 2 e as três últimas linhas referem-se ao

indivíduo 3. Portanto, agora é possível obter para cada indivíduo, as variâncias e covariâncias entre quaisquer pares de idades entre 11 e 52 meses. Tal abordagem pode ser vista em KIRKPATRICK et al. (1990). A matriz de coeficientes da função de covariância será estimada pela expressão $\hat{\mathbf{K}} = f^{-1}\hat{\mathbf{G}}(f^{-1})'$ (Passo 4). Neste caso, admite-se que a matriz de variâncias e covariâncias genética aditiva entre as idades (G) é conhecida. O resultado seria uma matriz de coeficientes da função de covariância de dimensão 3, pelo fato de terem sido utilizados os três primeiros polinômios de Legendre.

No caso de dados amostrais não se conhece a matriz G. Assim, MEYER e HILL (1997) propuseram uma metodologia que permite que os coeficientes da função de covariância possam ser estimados diretamente a partir dos dados, utilizando a metodologia da máxima verossimilhança restrita (REML), por meio de uma re-parametrização do algoritmo de REML, em modelos multivariados. Sendo assim, fornecendo um valor inicial para as matrizes de coeficientes da função de covariância, que agora podem ser denominadas de matrizes de coeficientes de regressão aleatória, para os efeitos genético aditivo e de ambiente permanente, $\hat{\mathbf{K}}\mathbf{a}$ e $\hat{\mathbf{K}}\mathbf{p}$, respectivamente, o algoritmo fornece as estimativas para essas matrizes.

Para a representação dos dados amostrais da Tabela 1, foram ajustadas funções de covariância tanto para o efeito genético aditivo quanto para o efeito de ambiente permanente, onde ambas as funções utilizaram os três primeiros polinômios de Legendre, caracterizando uma função polinomial de segundo grau. Este modelo pode ser descrito por: $\mathbf{y}_{ij} = \mathbf{F}_{ij} + \sum_{m=0}^{2} b_m f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} a_{im} f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} g_{im} f_m(\mathbf{a}^*_{ij}) + e_{ij}$, em que $\mathbf{y}_{ij}$ é a **j**-ésima produção do **i**-ésimo indivíduo; $\mathbf{a}^*_{ij}$ é a idade na produção padronizada entre –1 a +1; f_m é o **m**-ésimo polinômio de Legendre; $\mathbf{F}_{ij}$ é o efeito fixo de local; b_m são os coeficientes de regressão para modelar a trajetória média comum a todos os indivíduos; a_{im} e g_{im} são os coeficientes de regressão aleatória dos efeitos genético

aditivo e de ambiente permanente do indivíduo **i**, respectivamente, e e_{ij} é o efeito do ambiente temporário.

Em notação matricial o Modelo é representado como, $\mathbf{y} = \mathbf{X}b + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \mathbf{e}$,

sendo **y** o vetor referente a **n** observações de produção em cada idade; **X** é a matriz de incidência de níveis do efeito fixo de local e das idades padronizadas entre –1 a +1, associadas aos polinômios de Legendre, que descrevem a trajetória média de todos os indivíduos; **ß** é o vetor de soluções para níveis do efeito fixo de local e das soluções da regressão média de todos indivíduos; **Z** e **W** são matrizes de covariáveis referentes às idades padronizadas, associadas aos polinômios de Legendre, em cada produção, referentes aos coeficientes de regressão aleatória dos efeitos aleatórios genético aditivo e de ambiente permanente para cada indivíduo, respectivamente; **a** e **p** são vetores contendo os coeficientes de regressão aleatória para cada indivíduo, para os efeitos genético aditivo e de ambiente permanente, respectivamente. O vetor **e** representa os efeitos aleatórios de ambiente temporário. As pressuposições da distribuição dos vetores **a**, **p** e **e**, são as seguintes:

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} \sim \mathbf{N(0, V)} \text{ com } \mathbf{V} = \begin{bmatrix}\mathbf{A} \otimes \mathbf{Ka} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I} \otimes \mathbf{Kp} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{bmatrix},$$

sendo **A,** a matriz dos coeficientes de parentesco entre indivíduos, de ordem igual ao número de indivíduos (**N**); $\hat{\mathbf{K}}\mathbf{a}$ é a matriz de covariância entre os coeficientes de regressão aleatória do efeito genético aditivo; $\hat{\mathbf{K}}\mathbf{p}$ é a matriz de covariância entre os coeficientes de regressão aleatória que descrevem o efeito de ambiente permanente; **I** é uma matriz identidade, de ordem igual ao número total de observações (**n**) e **R** é uma matriz diagonal de variância residual associada a cada observação.

Para obtenção das matrizes de coeficientes de regressão aleatória para os efeitos genético aditivo e de ambiente permanente e posterior resolução das equações de modelos mistos, as seguintes matrizes precisam ser definidas:

A matriz de parentesco **A**, que define a relação de parentesco entre os três indivíduos avaliados da Tabela 1, será definida como:

$$\mathbf{A} = \begin{bmatrix} \mathbf{1} & \mathbf{0,25} & \mathbf{0} \\ \mathbf{0,25} & \mathbf{1} & \mathbf{0,25} \\ \mathbf{0} & \mathbf{0,25} & \mathbf{1} \end{bmatrix}.$$

O vetor de observações é representado como:

$$\mathbf{y} = \begin{bmatrix} \mathbf{15,1} \\ \mathbf{25,4} \\ \mathbf{42,1} \\ \mathbf{13,8} \\ \mathbf{45,4} \\ \mathbf{16,2} \\ \mathbf{23,9} \\ \mathbf{41,8} \end{bmatrix}$$

A matriz **X** contém os efeitos fixos (neste exemplo apenas o efeito fixo de local) e a variável regressora que gera a trajetória comum a todos os indivíduos (curva fixa). Nesta matriz, o efeito fixo de local, está sob a restrição de que a soma das soluções de locais é igual a zero.

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{0,7071} & \mathbf{-1,1649} & \mathbf{1,3553} \\ \mathbf{1} & \mathbf{0,7071} & \mathbf{0,2688} & \mathbf{-0,6763} \\ \mathbf{1} & \mathbf{0,7071} & \mathbf{0,9857} & \mathbf{0,7458} \\ \mathbf{-1} & \mathbf{0,7071} & \mathbf{-1,2247} & \mathbf{1,5811} \\ \mathbf{-1} & \mathbf{0,7071} & \mathbf{1,2247} & \mathbf{1,5811} \\ \mathbf{-1} & \mathbf{0,7071} & \mathbf{-1,0454} & \mathbf{0,9377} \\ \mathbf{-1} & \mathbf{0,7071} & \mathbf{0,2090} & \mathbf{-0,7214} \\ \mathbf{-1} & \mathbf{0,7071} & \mathbf{0,9259} & \mathbf{0,5652} \end{bmatrix}$$

A primeira linha da matriz **X** se refere à observação 15,1 do vetor **y**, do indivíduo 5 na idade aos 12 meses.

Como os dados estão ordenados por indivíduo e idade, a matriz **Z** (matriz da variável regressora idade do indivíduo para o efeito genético aditivo) será uma matriz bloco diagonal, onde cada bloco se refere a um indivíduo.

$$\mathbf{Z} = \begin{bmatrix} \mathbf{0,7071} & \mathbf{-1,1649} & \mathbf{1,3553} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0,7071} & \mathbf{0,2688} & \mathbf{-0,6763} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0,7071} & \mathbf{0,9857} & \mathbf{0,7458} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7071} & \mathbf{-1,2247} & \mathbf{1,5811} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7071} & \mathbf{1,2247} & \mathbf{1,5811} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7071} & \mathbf{-1,0454} & \mathbf{0,9377} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7071} & \mathbf{0,2090} & \mathbf{-0,7214} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7071} & \mathbf{0,9259} & \mathbf{0,5652} \end{bmatrix}$$

A matriz **W** (matriz da variável regressora idade do indivíduo para o efeito de ambiente permanente) é igual à matriz **Z**. Além disso, vamos admitir neste exemplo, que a variância do efeito de ambiente temporário tenha sido a mesma para todas as idades (homogeneidade de variâncias), sendo considerada 2,2 cm^2. Assim, **R** (matriz de variância do ambiente temporário) será uma matriz diagonal de dimensão igual a 8.

$$\mathbf{R} = \begin{bmatrix} \mathbf{2,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{2,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2,2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2,2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2,2} \end{bmatrix}.$$

Utilizando o algoritmo de MEYER e HILL (1997), vamos admitir que foram estimadas as seguinte matrizes:

$$\hat{\mathbf{K}}_a = \begin{bmatrix} \mathbf{2,2367} & \mathbf{0,5571} & \mathbf{-\,0,1227} \\ \mathbf{0,5571} & \mathbf{0,2939} & \mathbf{0,0027} \\ \mathbf{-\,0,1227} & \mathbf{0,0027} & \mathbf{0,0328} \end{bmatrix} \text{ e } \hat{\mathbf{K}}_p = \begin{bmatrix} \mathbf{1,4140} & \mathbf{0,1405} & \mathbf{-\,0,3454} \\ \mathbf{0,1405} & \mathbf{0,5960} & \mathbf{0,1328} \\ \mathbf{-\,0,3454} & \mathbf{0,1328} & \mathbf{0,4869} \end{bmatrix}.$$

Para a obtenção das funções que irão fornecer o valor genético para cada indivíduo, é necessária a utilização das equações de modelos mistos. As equações de modelos mistos, em um modelo de regressão aleatória podem ser assim definidas:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} & \mathbf{X'R^{-1}W} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + A^{-1} \otimes K_a^{-1}} & \mathbf{Z'R^{-1}W} \\ \mathbf{W'R^{-1X}} & \mathbf{W'R^{-1}Z} & \mathbf{W'R^{-1}W + I \otimes K_p^{-1}} \end{bmatrix} \begin{bmatrix} b \\ a \\ d \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \\ \mathbf{W'R^{-1}y} \end{bmatrix}$$

As soluções das equações de modelos mistos fornecem o valor da estimativa do efeito fixo de local e as soluções de regressão aleatória da parte fixa (vetor β), soluções de regressão aleatória do efeito genético aditivo (vetor α) e de ambiente permanente (vetor δ) para cada indivíduo, nas idades em que foram avaliados. Também é possível estimar valores genéticos para qualquer ponto na trajetória, no intervalo entre 11 e 52 meses (menor e maior idade avaliada).

A soluções encontradas para o efeito de local foram iguais a 0,278 e –0,278. Para a curva fixa as soluções para os parâmetros de intercepto, termo linear e termo quadrático foram iguais a 34,88, 13,17 e 4,38, respectivamente. Predizendo a variável DAP por meio da função de regressão fixa, utilizando as idades na escala padronizada, obtém-se a Figura 1:

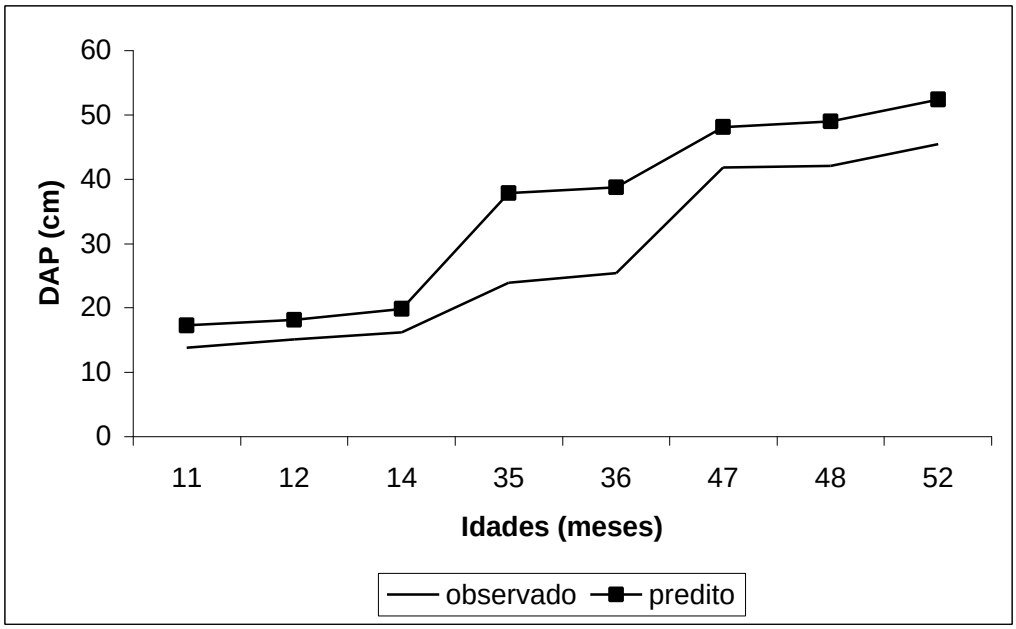

Figura 1: Valores observados e preditos da variável diâmetro à altura do peito em função das idades avaliadas, por meio da regressão fixa, comum a todos os indivíduos.

As soluções de regressão aleatória do efeito genético aditivo para os indivíduos 1, 2 e 3 estão apresentadas na Tabela 1. Como foi feito um ajuste quadrático, cada indivíduo apresentará três soluções, uma referente a cada parâmetro da função ajustada, ou seja, o intercepto, o termo linear e o termo quadrático.

Tabela 1 – Soluções de regressão aleatória (intercepto, linear e quadrático) do efeito genético aditivo para os indivíduos 1, 2 e 3.

Solução	Valor
a_1	0,062647
a_1	0,091115
a_1	0,009001
a_2	-0,17971
a_2	0,067594
a_2	0,019379
a_3	0,819711
a_3	0,266111
a_3	-0,01823

Para obtenção da estimativa da variância genética aditiva da variável em estudo para uma determinada idade, $\hat{s}^2_a$, basta multiplicar o vetor linha da idade padronizada e associada aos polinômios de Legendre (f_m) pela matriz de coeficientes de regressão aleatória do efeito genético aditivo ($\hat{\mathbf{K}}\mathbf{a}$) e pós-multiplicar o resultado pelo transposto do vetor f_m. Por exemplo, considerando as idades 11 e 52 meses, a variância genética em cada idade é obtida por meio da expressão:

$$\hat{s}^2_{a_{11}} = f_{11}\hat{\mathbf{K}}\mathbf{a}(f_{11})' =$$

$$= \begin{bmatrix}\mathbf{0,7071} & \mathbf{-1,1649} & \mathbf{1,3553}\end{bmatrix} \begin{bmatrix}\mathbf{2,2367} & \mathbf{0,5571} & \mathbf{-0,1227} \\ \mathbf{0,5571} & \mathbf{0,2939} & \mathbf{0,0027} \\ \mathbf{-0,1227} & \mathbf{0,0027} & \mathbf{0,0328}\end{bmatrix} \begin{bmatrix}\mathbf{0,7071} \\ \mathbf{-1,1649} \\ \mathbf{1,3553}\end{bmatrix} =$$

$$= \mathbf{0,415cm^2}$$

$$\hat{s}^2_{a_{52}} = f_{52}\hat{\mathbf{K}}\mathbf{a}(f_{52})' =$$

$$= \begin{bmatrix}\mathbf{0,7071} & \mathbf{0,9259} & \mathbf{0,5652}\end{bmatrix} \begin{bmatrix}\mathbf{2,2367} & \mathbf{0,5571} & \mathbf{-0,1227} \\ \mathbf{0,5571} & \mathbf{0,2939} & \mathbf{0,0027} \\ \mathbf{-0,1227} & \mathbf{0,0027} & \mathbf{0,0328}\end{bmatrix} \begin{bmatrix}\mathbf{0,7071} \\ \mathbf{0,9259} \\ \mathbf{0,5652}\end{bmatrix} =$$

$$= \mathbf{2,015cm^2}$$

A covariância genética aditiva entre as idades 11 e 52 meses é obtida como:

$$\hat{s}_{a_{11;52}} = f_{11}\hat{\mathbf{K}}\mathbf{a}(f_{52})'$$

$$\hat{s}_{a_{11,52}} = f_{11}\hat{\mathbf{K}}\mathbf{a}(f_{52})' =$$

$$= \begin{bmatrix}\mathbf{0,7071} & \mathbf{-1,1649} & \mathbf{1,3553}\end{bmatrix} \begin{bmatrix}\mathbf{2,2367} & \mathbf{0,5571} & \mathbf{-0,1227} \\ \mathbf{0,5571} & \mathbf{0,2939} & \mathbf{0,0027} \\ \mathbf{-0,1227} & \mathbf{0,0027} & \mathbf{0,0328}\end{bmatrix} \begin{bmatrix}\mathbf{0,7071} \\ \mathbf{0,9259} \\ \mathbf{0,5652}\end{bmatrix} = \mathbf{0,567cm}$$

Para obtenção das variâncias e covariância para o efeito de ambiente permanente para duas idades quaisquer, basta substituir a matriz $\hat{\mathbf{K}}\mathbf{a}$ pela matriz $\hat{\mathbf{K}}\mathbf{p}$ (matriz de coeficientes de regressão aleatória para o efeito de ambiente permanente). Assim as variâncias para este efeito sobre a DAP aos 11 e 52 meses de idade foram 1,097 e 1,420 cm^2, respectivamente. A covariância entre elas foi igual 0,023 cm.

Com base nas estimativas de variância e covariância entre as idade, pode-se então obter as estimativas de herdabilidade ($\hat{\mathbf{h}}^2$) e de correlação genética ($\mathbf{r_a}$) para DAP aos 11 e 52 meses de idade como:

$$\hat{\mathbf{h}}^2_{11} = \frac{\hat{S}^2_{a_{m1}}}{\hat{S}^2_{a_{m1}} + \hat{S}^2_{p_{m1}} + \hat{S}^2_{e_{m1}}} = \frac{\mathbf{0,415}}{\mathbf{0,415} + \mathbf{1,097} + \mathbf{2,200}} = 0,11$$

$$\hat{\mathbf{h}}^2_{52} = \frac{\mathbf{2,015}}{\mathbf{2,015} + \mathbf{1,420} + \mathbf{2,200}} = 0,36$$

$$\mathbf{r}_{a_{11,52}} = \frac{S_{a_{im1,m2j}}}{\sqrt{S^2_{m1} S^2_{m2}}} = \frac{\mathbf{0,567}}{\sqrt{\mathbf{0,415} + \mathbf{2,015}}} = 0,62$$

Utilizando-se combinações lineares de uma determinada idade padronizada e associada ao polinômio de Legendre (**m**) com as soluções de regressão aleatória do efeito genético aditivo para cada indivíduo (**i**), é possível obter o valor genético para cada indivíduo nesta idade, por meio da seguinte expressão $\mathbf{VG_i} = \left[f_m(\mathbf{a}^*_{ij})\right] a_{im}$. Neste caso, para a obtenção do valor genético da DAP do indivíduo 2 aos 36 meses de idade, mesmo que o indivíduo 2 não tenha sido mensurado nesta idade, tem-se:

$$\mathbf{VG}_{2,36} = \begin{bmatrix}\mathbf{0,7071} & \mathbf{0,2688} & \mathbf{-0,6763}\end{bmatrix} \begin{bmatrix}\mathbf{-0,17971} \\ \mathbf{0,067594} \\ \mathbf{0,019379}\end{bmatrix} = -0,122.$$

Capítulo 2

Simulação de dados

A técnica de simulação de dados para a estrutura de modelos mistos tem sido descrita por VAN VLECK (1994), VERNEQUE (1994), SCHAEFFER (1997); FERREIRA e BEARZOTI (2003), entre outros, utilizando a decomposição de Cholesky das matrizes de covariâncias dos efeitos aleatórios do modelo. Em uma situação geral, se **K** representa uma matriz de covariância qualquer, sendo **K=L . L´**, onde **L** é a decomposição de Cholesky da matriz **K** e fazendo a transformação **y=** m **+ L . Z**, com **Z** uma matriz de variáveis aleatórias normais padronizadas (**Z ~ N(0,1)**), então **E**(**y**) = m e **V**(**y**) = **0** + **L . L´ = K**. Assim **y** possui distribuição normal com média m e variância **K** **(Y~ N(**m**, K))**.

No caso de dados longitudinais, admitindo que será criado um conjunto de dados de **N** indivíduos com cada indivíduo contendo registros em **i** idades diferentes, o modelo linear misto é definido como: **y** = **X**b + **Za** + **Wp** + e em que **y** é um vetor de observações dos indivíduos; b é um vetor de parâmetros de efeitos fixos e dos parâmetros da curva geral para todos os indivíduos; **X** é uma matriz de incidência dos níveis de efeitos e com as variáveis regressoras, representadas pelas idades padronizadas associadas aos Polinômios de Legendre; **a** e **p** são vetores aleatórios de coeficientes de regressão aleatória do efeito genético aditivo e de regressão aleatória do efeito de ambiente permanente, respectivamente; **Z** e **W** são matrizes que associam as variáveis regressoras aos vetores **a** e **p** de efeitos aleatórios, respectivamente. O vetor e é um vetor de efeito aleatório de ambiente temporário.

Assumindo que os vetores **y**, **a**, **p** e e apresentam distribuição normal, a esperança do vetor aleatório **a** é **E(a)** = Æ e a variância é **V(a)** = **A** Ä **Ka** = **G** , onde **Ka** é uma matriz de covariâncias entre os coeficientes de regressão aleatória do efeito genético aditivo e **A** é a matriz que indica o grau de parentesco entre os indivíduos, de dimensão igual ao

número total de indivíduos (**N**). O vetor **p** possui **E(a)** = Æ e a variância é **V(p)** = **I** Ä **Kp** = **P**, onde **Kp** é uma matriz de covariâncias entre os coeficientes de regressão aleatória do efeito de ambiente permanente e **I** uma matriz identidade de dimensão igual ao número de indivíduos com observações (**n**). Por último, e possui média **E(**e**)** = Æ e variância **V(**e**)** = **I** s_e^2 = **R**, onde s_e^2 é a variância do efeito de ambiente temporário. Conseqüentemente, a esperança e a variância do vetor de dados observados **y**, são **E(y)** = **X**b e **V(y)** = **ZGZ'**+**WPW´**+**R**, respectivamente. Ao vetor b estão associados os efeitos fixos e a curva de regressão geral para todos os indivíduos. Se $\mathbf{L_a}$ e $\mathbf{L_p}$ são matrizes triangulares inferiores obtidas pelas decomposições de Cholesky das matrizes de covariâncias dos coeficientes de regressão aleatória dos efeitos genético aditivo e de ambiente permanente, respectivamente, e **A** é uma matriz de numeradores de coeficiente de parentesco entre os indivíduos de ordem N, onde $\mathbf{T_A}$ é também uma matriz triangular inferior obtida pela decomposição de Cholesky dessa matriz de parentesco, de mesma ordem. Então o vetor **y** (fenótipos) contendo as "**i**" características (idades) é definido como:

y = m **+$\mathbf{T_A}$. Za**. $\mathbf{L_a}$ **+ Zp.** $\mathbf{L_p}$ + **e**, ou seja **y**=média + valor genético aditivo+ valor de ambiente permanente + valor de ambiente temporário. Assim seria criado um conjunto de **i** idades, com distribuição normal, onde cada efeito aleatório possui a sua estrutura própria de covariância, e estes mesmos efeitos aleatórios seriam independentes entre si.

A simulação dos dados foi realizada por meio de um modelo polinomial de segundo grau, utilizando os polinômios ortogonais de Legendre para descrever tanto a trajetória fixa, quanto à dos efeitos aleatórios que compõem o modelo. Os dados foram simulados de forma a representar uma estrutura de indivíduos contendo informação do fenótipo Diâmetro à Altura do Peito (DAP) de cada um dos indivíduos, às idades de 12, 30, 48, 66 e 84 meses foi representado. O modelo foi representado como:

$$\mathbf{y_{ijk}} = \mathbf{L_j} + (\mathbf{b_0} + \mathbf{b_1 A} + \mathbf{b_2 A^2}) + (\mathbf{a_{i0}} + \mathbf{a_{i1} A} + \mathbf{a_{i2} A^2}) + (\mathbf{p_{i0}} + \mathbf{p_{i1} A} + \mathbf{p_{i2} A^2}) + \mathbf{e_{ijk}}$$

onde

$\mathbf{L_j}$ é o efeito fixo de local, com **j**=1, 2 e 3; $\mathbf{b_o}$, $\mathbf{b_1}$ e $\mathbf{b_2}$ são coeficientes de regressão que descrevem a curva fixa geral da produção em função das idades padronizadas entre –1 a +1 associadas aos três primeiros polinômios de Legendre; $\mathbf{a_{i0}}$, $\mathbf{a_{i1}}$ e $\mathbf{a_{i2}}$ são os coeficientes de regressão aleatória relacionados ao efeito genético aditivo para o indivíduo **i**, assumidos como tendo distribuição multinormal com vetor de média zero e matriz de covariâncias **A**Ä**Ka**; $\mathbf{p_{i0}}$, $\mathbf{p_{i1}}$ e $\mathbf{p_{i2}}$ são os coeficientes de regressão aleatória relacionados ao efeito de ambiente permanente para o indivíduo **i**, assumidos como tendo distribuição multinormal com vetor de média zero e matriz de covariâncias **I**Ä**Kp**; $\mathbf{e_{ijk}}$ é o efeito residual temporário assumido tendo distribuição normal com média zero e variância s^2_e=2,2 unidades^2, para dados gerados de modo a apresentar homogeneidade de variâncias para esse efeito, e variâncias iguais a 2,2 e 4,4 unidades^2, para as três primeiras e duas últimas idades, respectivamente, para dados gerados de modo a apresentar heterogeneidade de variâncias.

Os dados de produção foram simulados inicialmente, por meio de geração aleatória de matrizes **a** e **p**, que representam as soluções de regressão aleatória dos efeitos genético direto e de ambiente permanente, respectivamente. Para tanto, foram obtidas as matrizes **Ka** e **Kp**, que representam as matrizes de covariâncias dos coeficientes de regressão aleatória dos efeitos genético aditivo e de ambiente permanente, respectivamente. As matrizes **Ka** e **Kp** apresentaram ordem três, que corresponde a um ajuste polinomial de segundo grau.

Para a obtenção da matriz **Ka**, foi determinada inicialmente uma matriz de covariâncias genéticas aditivas entre as idades ($\mathbf{G_0}$) de ordem cinco, de forma a apresentar variâncias maiores nas idades finais e uma estrutura de correlação maior entre idades mais próximas ($\mathbf{R_G}$). A matriz $\mathbf{G_0}$ foi obtida de forma que uma função de covariância, que empregou os três primeiros polinômios de Legendre, fosse suficiente para recriar $\mathbf{G_0}$ sem falta de ajuste.

$$\mathbf{G_0} = \begin{bmatrix} \mathbf{0,3915} & \mathbf{0,4828} & \mathbf{0,5289} & \mathbf{0,5296} & \mathbf{0,4850} \\ \mathbf{0,4828} & \mathbf{0,7824} & \mathbf{0,9693} & \mathbf{1,0437} & \mathbf{1,0055} \\ \mathbf{0,5289} & \mathbf{0,9693} & \mathbf{1,2761} & \mathbf{1,4492} & \mathbf{1,4885} \\ \mathbf{0,5296} & \mathbf{1,0437} & \mathbf{1,4492} & \mathbf{1,7460} & \mathbf{1,9341} \\ \mathbf{0,4850} & \mathbf{1,0055} & \mathbf{1,4885} & \mathbf{1,9341} & \mathbf{2,3423} \end{bmatrix}$$

$$\mathbf{R_G} = \begin{bmatrix} \mathbf{1} & \mathbf{0,8724} & \mathbf{0,7482} & \mathbf{0,6406} & \mathbf{0,5065} \\ \mathbf{0,8724} & \mathbf{1} & \mathbf{0,9701} & \mathbf{0,8930} & \mathbf{0,7427} \\ \mathbf{0,7482} & \mathbf{0,9701} & \mathbf{1} & \mathbf{0,9709} & \mathbf{0,8610} \\ \mathbf{0,6406} & \mathbf{0,8930} & \mathbf{0,9709} & \mathbf{1} & \mathbf{0,9564} \\ \mathbf{0,5065} & \mathbf{0,7427} & \mathbf{0,8610} & \mathbf{0,9564} & \mathbf{1} \end{bmatrix}$$

A matriz $\mathbf{G_0}$ foi determinada de modo que uma função de covariância utilizando os três primeiros polinômios de Legendre (função quadrática) fosse suficiente para descrever a estrutura de covariância entre as idades. Assim:

$$\mathbf{G_0^*} = f\ \mathbf{Ka}\ f',$$

onde $\mathbf{G_0^*}$ é a matriz de covariância genética aditiva estimada pela função de covariância, f é a matriz de idades padronizadas (**M**) associadas aos polinômios de Legendre (L), ou seja, $f = \mathbf{M}\ L$, em que

$$\mathbf{M} = \begin{bmatrix} \mathbf{1} & \mathbf{-1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-0,5} & \mathbf{0,25} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0,5} & \mathbf{0,25} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix} \quad \text{e} \quad L = \begin{bmatrix} \mathbf{0,7071} & \mathbf{0} & \mathbf{0,7906} \\ \mathbf{0} & \mathbf{1,2247} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2,3717} \end{bmatrix}$$

A matriz **Ka**, matriz de covariâncias dos coeficientes de regressão aleatória do efeito genético aditivo, foi obtida por meio da expressão:

$$\mathbf{Ka} = (f'f)^{-1} f'\mathbf{G}_0 f (f'f)^{-1}$$

Para a obtenção da matriz **Kp**, matriz de covariâncias de coeficientes de regressão aleatória do efeito de ambiente permanente, a matriz de covariâncias do efeito de ambiente permanente entre as idades ($\mathbf{P_0}$) foi determinada de modo a apresentar uma estrutura de covariância possível de ser recriada por uma função de covariância utilizando os três primeiros polinômios de Legendre, ou seja, $\mathbf{P_0^*} = f\,\mathbf{Kp}\,f'$, onde $\mathbf{P_0^*}$ é a matriz de covariâncias do efeito de ambiente permanente entre as idades, estimada pela função de covariância. De maneira análoga à matriz **Ka**, a matriz **Kp** foi obtida por meio da expressão:

$$\mathbf{Kp} = (f'f)^{-1} f'\mathbf{P_0} f (f'f)^{-1}$$

$$\mathbf{Kp} = \begin{bmatrix} \mathbf{1,4140} & \mathbf{0,1405} & \mathbf{-0,3454} \\ \mathbf{0,1405} & \mathbf{0,5960} & \mathbf{0,1328} \\ \mathbf{-0,3454} & \mathbf{0,1328} & \mathbf{0,4869} \end{bmatrix}$$

em que,

$$\mathbf{P_0} = \begin{bmatrix} \mathbf{1,2881} & \mathbf{0,3849} & \mathbf{-0,0878} & \mathbf{-0,1302} & \mathbf{0,2578} \\ \mathbf{0,3849} & \mathbf{0,9565} & \mathbf{1,0279} & \mathbf{0,5991} & \mathbf{-0,3299} \\ \mathbf{-0,0878} & \mathbf{1,0279} & \mathbf{1,3975} & \mathbf{1,0211} & \mathbf{-0,1015} \\ \mathbf{-0,1302} & \mathbf{0,5991} & \mathbf{1,0211} & \mathbf{1,1357} & \mathbf{0,9429} \\ \mathbf{0,2578} & \mathbf{-0,3299} & \mathbf{-0,1015} & \mathbf{0,9429} & \mathbf{2,8035} \end{bmatrix}$$

A estrutura de pedigree responsável pela geração da matriz de parentesco entre indivíduos, constituiu-se de 30 progenitores masculinos, acasalados cada um com 3

progenitores femininos diferentes, gerando em cada acasalamento 10 proles, totalizando em 1020 indivíduos (**N**), sendo 900 indivíduos com informação (**n**).

Para gerar o vetor de coeficientes de regressão aleatória do efeito genético aditivo para cada indivíduo (**a**), foi gerada a matriz "**g**", contendo **N** linhas e três colunas, onde cada coluna constituiu uma variável aleatória normal padronizada. Essa mesma matriz **g** foi pré-multiplicada pela decomposição de Cholesky da matriz **Ka**, **D(Ka)**, de forma a gerar um intercepto, um termo linear e um termo quadrático para cada individuo que constituiu a amostra. Posteriormente, multiplicou-se o produto dessas matrizes (**D(Ka) g**) pela decomposição de Cholesky da matriz de coeficientes de parentesco entre os indivíduos (**D(A)**). Em seguida foi gerada uma matriz **Z**, que corresponde à matriz de covariáveis regressoras, composta pelo produto direto de uma matriz identidade de dimensão **N** com uma matriz f, de ordem 5x3, representando as cinco idades padronizadas entre –1 a 1 e associadas aos três primeiros Polinômios de Legendre nas linhas, e o intercepto, termos linear e quadrático para cada idade na coluna ($\mathbf{Z=I_{(N)}} \otimes f$). Para os indivíduos sem informação, que são os progenitores, ao invés de covariáveis regressoras em **Z**, as linhas e colunas desses indivíduos foram ocupadas por zeros. Em resumo, o termo **Za** do modelo linear misto, foi gerado da seguinte forma:

$$\mathbf{Za = \{[I_{(N)} \otimes f\,] [D(Ka)\ g\ D(A)]\} = f_a\ a}$$

Em que;

$$\mathbf{D(Ka)} = \begin{bmatrix} \mathbf{1,4956} & \mathbf{0} & \mathbf{0} \\ \mathbf{0,3725} & \mathbf{0,3939} & \mathbf{0} \\ \mathbf{-\ 0,0821} & \mathbf{0,0845} & \mathbf{0,1377} \end{bmatrix}.$$

Na geração do vetor de coeficientes de regressão aleatória do efeito de ambiente permanente (**p**), para os indivíduos com informação, foi gerada a matriz **pe**, contendo **n** linhas e três colunas, onde cada coluna também constituiu uma variável aleatória normal padronizada. Essa mesma matriz **pe** foi pré-multiplicada pela decomposição de Cholesky

da matriz **Kp**, **D(Kp)**. Posteriormente, multiplicou-se o produto dessas matrizes (**D(Kp).pe**) por uma matriz **W** , onde **W** foi obtida pelo produto direto de uma matriz identidade de ordem **n** com a matriz f **, W=$I_{(n)}$** Ä f , resultando em **Wp** do modelo linear misto, ou seja,

Wp = {[I(n) Ä f **] [D(Kp) pe]} =** f_p **p**,

onde,

$$\mathbf{D(Kp)} = \begin{bmatrix} \mathbf{1,1891} & \mathbf{0} & \mathbf{0} \\ \mathbf{0,1182} & \mathbf{0,7629} & \mathbf{0} \\ \mathbf{-0,2905} & \mathbf{0,2190} & \mathbf{0,5954} \end{bmatrix}.$$

Para o efeito de ambiente temporário foi gerado um vetor contendo **n** linhas de uma variável com distribuição normal padronizada, pré-multiplicado pelo desvio padrão residual desejado para a amostra.

Os parâmetros fixos da curva geral dos indivíduos foram estabelecidos de modo a determinar uma curva com a seguinte característica, apresentada na figura 1.

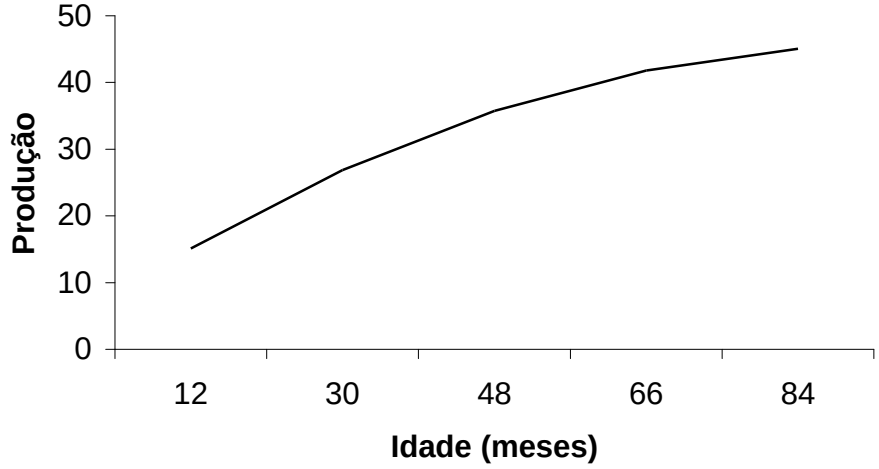

FIGURA 1- Representação gráfica da regressão das idades padronizadas em função da produção dos indivíduos, utilizando os três primeiros polinômios ortogonais de Legendre.

O vetor b contendo os parâmetros da curva fixa foram obtidos pela expressão $b = (f' f)^{-1}(f'\mathbf{Y})$, onde **Y=[15,14; 26,82; 35,72; 41,84; 45,1]**; resultando em $b' = ((f' f)^{-1}(f'\mathbf{y}))'$ =**[47,8832 12,2381 –2,3636]**. O vetor b foi pré-multiplicado pelo produto direto entre um vetor com 900 linhas contendo somente com números iguais à um e a matriz f ; totalizando em um vetor bloco coluna de dimensão 4.500x3.

O efeito fixo de local foi criado de forma que, cada combinação de progenitores gerasse 10 proles distribuídas em três locais diferentes. Não sendo esperadas diferenças significativas entre locais.

Assim os dados de produção de 900 indivíduos gerados pela combinação entre 10 progenitores masculinos com 30 progenitores femininos, contendo 5 informações para cada indivíduo, referentes às idades aos 12, 30, 48, 66 e 84 meses de idade, totalizando 4500 observações obtidas por:

$$\mathbf{y} = [\mathbf{I}_{(900)} \otimes f\,]\, b + \{[\mathbf{I}_{(N)} \otimes f\,]\,[\mathbf{D(Ka)}\, g\, \mathbf{D(A)}]\} + \{[\mathbf{I}_{(n)} \otimes f\,]\,[\mathbf{D(Kp)}\, pe]\} + \mathbf{I}_{(n)}\, s_e^2,$$

$$\mathbf{y} = \mathbf{X}b + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + e,$$

tal que,

$$\mathbf{y}_{ij} = \mathbf{F}_{ij} + \sum_{m=0}^{2} b_m f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} a_{im} f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} g_{im} f_m(\mathbf{a}^*_{ij}) + e_{ij},$$

em que $\mathbf{y}_{ij}$ é a **j**-ésima produção do **i**-ésimo indivíduo, $\mathbf{a}^*_{ij}$ é a idade na produção padronizada entre –1 a +1, $f_m(\mathbf{a}^*_{ij})$ é o **m**-ésimo polinômio de Legendre, $\mathbf{F}_{ij}$ é o efeito fixo de local, b_m são os coeficientes de regressão para modelar a trajetória média de todos os indivíduos, a_{im} e g_{im} são os coeficientes de regressão aleatória dos efeitos genético aditivo e de ambiente permanente do indivíduo e e_{ij} é o efeito do ambiente temporário.

As matrizes de covariância entre os coeficientes de regressão aleatória para o efeito genético aditivo e de ambiente permanente (Ka e Kp), provenientes dos dados gerados, foram iguais a:

$$\mathbf{Ka} = \begin{bmatrix} \mathbf{2,24679} & \mathbf{0,55531} & \mathbf{-\,0,10496} \\ \mathbf{0,55531} & \mathbf{0,29601} & \mathbf{0,00495} \\ \mathbf{-\,0,10496} & \mathbf{0,00495} & \mathbf{0,03081} \end{bmatrix} \text{ e } \mathbf{Kp} = \begin{bmatrix} \mathbf{1,42628} & \mathbf{0,17282} & \mathbf{-\,0,34039} \\ \mathbf{0,17282} & \mathbf{0,60670} & \mathbf{0,14482} \\ \mathbf{-\,0,34039} & \mathbf{0,14482} & \mathbf{0,48634} \end{bmatrix},$$

respectivamente. A variância do efeito de ambiente temporário gerada pelos dados foi igual a 2,17 unidades^2, quando esse efeito foi criado homogêneo e iguais a 2,15 e 4,38 unidades^2 , quando o mesmo foi gerado de forma a apresentar-se heterogêneo.

REFERÊNCIAS BIBLIOGRÁFICAS

VERNEQUE, R. S. Procedimentos numéricos e estimação de componentes de covariância pelo método da máxima verossimilhança restrita, na análise multivariada, aplicados a modelos de melhoramento animal. Piracicaba. 1994. 155p. (Doutorado-ESALQ/USP)

VAN VLECK. L. D. Algoritms for siulations of animal models with múltiple traits and non-additive genetic effects. Journal Sr., Nebraska Agric. Res. Div., 74(9): 3174-3182, 1994

FERREIRA E. B., BEARZOTI, E. Comparação de métodos no ajustamento de curvas de lactação de bovinos por meio de simulação, Ciênc. Agrotec., Lavras. V.27, n.4, p.865-872, jul./ago., 2003

SCHAEFFER, L. R. Random regression models.In: SCHAEFFER, L. R.; VAN DER WERF, J. **Course notes**. Ontario: University of Guelph, 1997. 104 p.

Capítulo 3

Estimação de parâmetros genéticos e componentes de variância, utilizando modelos de regressão aleatória sob diferentes pressuposições em dados simulados

Resumo: Dados longitudinais de características de crescimento, em um teste de progênie obtidos por meio de simulação, foram analisados em diferentes modelos de regressão aleatória, para avaliar o comportamento das estimativas dos componentes de variância e dos parâmetros genéticos. Foram assumidas diferentes pressuposições para as variâncias do efeito ambiental. Modelos de regressão aleatória que ajustaram ou não uma função de covariância para o efeito de ambiente permanente foram utilizados, considerando ou não classes heterogêneas de variância do efeito de ambiente temporário, gerando quatro diferentes modelos. Em todos os modelos foram utilizados os polinômios de Legendre até a terceira ordem (ajuste de segundo grau) para os efeitos genético aditivo e de ambiente permanente, quando ajustado. Para a variância do efeito de ambiente temporário, foram considerados dois valores diferentes. De acordo com o teste da razão de verossimilhança e do Critério de Akaike, o modelo que mais se aproximou da situação real, foi o que considerou tanto a variância do efeito de ambiente permanente como a do efeito de ambiente temporário. Além disso, modelos que consideraram a heterogeneidade de variância do efeito de ambiente temporário foram mais adequados do que modelos que consideraram a variância do efeito de ambiente permanente. Apesar de os resultados obtidos pelo Modelo 2 terem se apresentado mais próximos do Modelo 1, eleito o melhor modelo, os testes indicaram o Modelo 3 como o segundo melhor modelo. Assim, enquanto o Modelo 3 proporcionou maior precisão, o Modelo 2 apresentou maior acurácia.

Palavras-chave: dados longitudinais, heterogeneidade de variâncias, parâmetros genéticos, características de crescimento.

Chapter 3

Evaluation of the heterogeneity of variance effect using random regression models with simulated data

Abstract: Longitudinal data of growth traits, in a simulated progeny test, were analyzed by different random regression models to evaluate the estimates of variance components and genetic parameters. Different assumptions for the environmental variance effects were assumed. Random regression models with and without a covariance function associated to the permanent environmental effect were used, considering or not heterogeneous classes of temporary environmental effect variance, giving four different models. In all models, Legendre's polynomials of third order were used (second degree fit) for additive genetic and permanent environmental effects, when fitted. For temporary environmental effect variance, we considered two different values. According to the likelihood ratio test and Akaike's criterion, the model that was more similar to the real situation, was that considered both permanent environmental effect and the temporary one. Moreover, models that accounted for the heterogeneity variance of temporary environmental effect were more adequate than models that accounted for the permanent environmental effect variance. Although the results obtained by Model 2 were more closer to Model 1, selected as the best one, the tests indicated Model 3 as the second better model. So, while Model 3 was more precise, Model 2 was more accurate.

Key words: longitudinal data, heterogeneity of variance, genetic parameters, growth traits.

INTRODUÇÃO

A realização de medidas repetidas no tempo em um mesmo indivíduo é bastante comum no melhoramento de espécies vegetais perenes. Dados oriundos de medidas repetidas no mesmo indivíduo apresentam uma estrutura peculiar, onde medidas mais próximas são normalmente mais fortemente correlacionadas do que medidas mais distantes no tempo. Sendo assim, ao se adotar um modelo de repetibilidade para analisar tais dados, assumir que as correlações entre medidas repetidas são iguais à unidade, e, portanto, que todas as covariâncias são de mesma magnitude, nem sempre é verdade. Para a utilização de um modelo de características múltiplas, que seria teoricamente o mais correto, é necessário que os vários níveis em que as variáveis são mensuradas, sejam os mesmos para todos os indivíduos e, além disso, quando se têm muitos níveis a serem avaliados, este modelo se torna muito parametrizado, tornando as análises computacionalmente limitadas. Um exemplo de medições tomadas ao longo do tempo no mesmo indivíduo, seria a mensuração da variável diâmetro à altura do peito, em eucalipto.

Uma alternativa para a análise de dados de medidas repetidas que vem ganhando a atenção dos pesquisadores da área de melhoramento genético, é o uso de modelos de regressão aleatória. A utilização de modelos de regressão aleatória permite obter diferentes curvas de valores genéticos para cada indivíduo de modo a considerar as mudanças nas variâncias genéticas e residuais durante o tempo, e também a predição de valores genéticos dos indivíduos nas diferentes idades. O uso de modelos de regressão aleatória para modelar dados de medidas repetidas para a avaliação genética foi introduzido por pesquisadores da área animal (SCHAEFFER e DEKKERS, 1994; JAMROZIK e SCHAEFFER, 1997 e JAMROZIK et al., 1997, MEYER, 2004).

Segundo SCHAEFFER (2004), modelos de regressão aleatória poderiam ser perfeitamente aplicados ao crescimento de plantas, tais como culturas que crescem rapidamente ou árvores que crescem lentamente. Atualmente esta técnica tem sido muito utilizada em melhoramento animal, em diversas características e permitem ao pesquisador

estudar mudanças na variabilidade genética com o tempo e também a seleção de indivíduos para alterar o padrão de resposta sobre o tempo.

Particularmente, os modelos de regressão aleatória acomodam medidas repetidas para características que mudam gradualmente e continuamente no tempo e, além disso, não requerem pressuposições severas sobre a constância de variâncias e correlações (MEYER, 2000).

Uma vantagem do modelo de regressão aleatória é que não existe a necessidade de que os indivíduos apresentem mensurações em todas as idades e nem nas mesmas idades. Além disso, é possível incluir no estudo, indivíduos que apresentam apenas uma única mensuração. No entanto, as pressuposições assumidas a respeito do modelo podem levar a resultados imprecisos. Uma dessas pressuposições seria com relação à estrutura de variância residual, que inclui tanto a variância do efeito de ambiente permanente como a do efeito de ambiente temporário. De acordo com JENSEN (2001), em gado de leite, a variância residual geralmente é heterogênea durante a lactação e em partos diferentes. Uma das vantagens no uso de modelos de regressão aleatória está justamente na possibilidade de se modelar a estrutura de correlação quando se tem dados longitudinais.

Este estudo teve como objetivo verificar a importância de se considerar ou não a heterogeneidade de variâncias para o efeito de ambiente temporário, na utilização de modelos de regressão aleatória a dados que realmente apresentam esse comportamento, em diferentes situações:

a) Ajustando uma função polinomial de segundo grau tanto para o efeito genético aditivo como para o efeito de ambiente permanente;
b) Ajustando uma função polinomial de segundo grau para o efeito genético aditivo e uma função ortogonal de primeiro grau para o efeito de ambiente permanente, onde esse ajuste de uma função ortogonal de primeiro grau equivale a considerar que o efeito de ambiente permanente é constante.

MATERIAL E MÉTODOS

Uma estrutura de pedigree foi gerada por meio do cruzamento entre 30 progenitores masculinos com três progenitores femininos diferentes cada um, originando dez proles diferentes em cada cruzamento, distribuídas em três locais diferentes. O efeito fixo de local foi gerado de forma a não apresentar diferenças estatísticas significativas. Para cada prole foram geradas informações de fenótipos em cinco idades diferentes, ou seja, aos 12, 30, 48, 66 e 84 meses, resultando em 120 progenitores (30 masculinos e 90 femininos), 1020 indivíduos no total, sendo que 900 indivíduos (somente os descendentes) possuíam informação de produção em cinco idades diferentes, totalizando 4500 registros de avaliação de Diâmetro à Altura do Peito (DAP) .

A trajetória da produção média, comum para todos os indivíduos, foi gerada de forma a apresentar uma curva de comportamento quadrático (Figura 1). O vetor b contendo os parâmetros da curva fixa foram obtidos pela expressão $b' = \mathbf{[(}f'f\mathbf{)}^{-1}\mathbf{(}f'\mathbf{y)]}'$, onde **Y = [15,14; 26,82; 35,72; 41,84; 45,1]** representa a média de produção aos 12, 30, 48, 66 e 84 meses, respectivamente, e a matriz f representa o produto entre a matriz de idades padronizadas (**M**) e a matriz contendo os três primeiros polinômios de Legendre (L), $f = \mathbf{M}.L$, em que

$$\mathbf{M}=\begin{bmatrix}\mathbf{1} & \mathbf{-1} & \mathbf{1}\\ \mathbf{1} & \mathbf{-0,5} & \mathbf{0,25}\\ \mathbf{1} & \mathbf{0} & \mathbf{0}\\ \mathbf{1} & \mathbf{0,5} & \mathbf{0,25}\\ \mathbf{1} & \mathbf{1} & \mathbf{1}\end{bmatrix};\ L=\begin{bmatrix}\mathbf{0,7071} & \mathbf{0} & \mathbf{0,7906}\\ \mathbf{0} & \mathbf{1,2247} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{2,3717}\end{bmatrix}\ \text{e}\ f=\begin{bmatrix}\mathbf{0,7071} & \mathbf{-1,2247} & \mathbf{1,5811}\\ \mathbf{0,7071} & \mathbf{-0,6123} & \mathbf{-0,1976}\\ \mathbf{0,7071} & \mathbf{0} & \mathbf{-0,7906}\\ \mathbf{0,7071} & \mathbf{0,6123} & \mathbf{-0,1976}\\ \mathbf{0,7071} & \mathbf{1,2247} & \mathbf{1,5811}\end{bmatrix},$$

resultando em $b^{\circ} = \left(\left(f'f\right)^{-1}\mathbf{(}f'\mathbf{y)}\right)^{\circ} =$ **[47,8832 12,2381 –2,3636]**, que é o vetor de soluções para a curva fixa (Figura 1), comum a todos os indivíduos envolvidos na análise e que representa os coeficientes intercepto, linear e quadrático, da equação, respectivamente.

O efeito fixo de local foi criado de forma que cada combinação de progenitores gerasse 10 proles distribuídas em três locais diferentes, não sendo esperadas diferenças significativas entre locais.

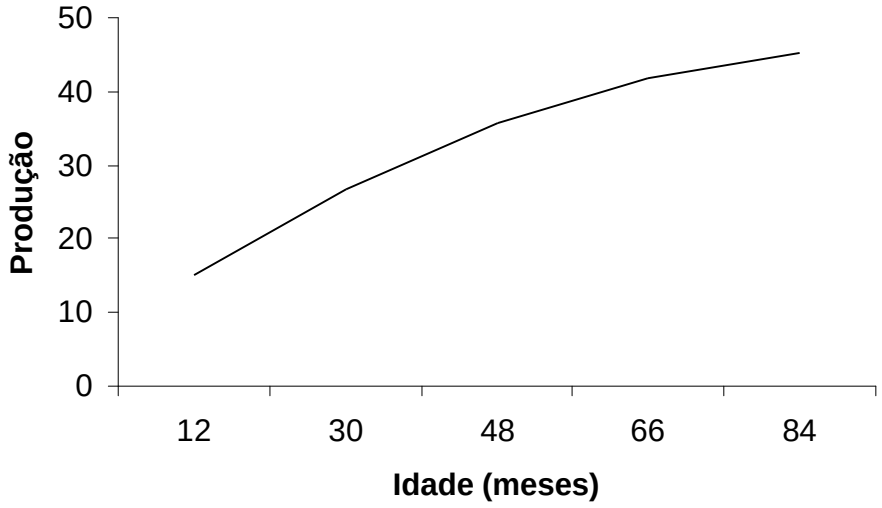

FIGURA 1- Representação gráfica da regressão da produção em função das idades dos indivíduos, utilizando os três primeiros polinômios ortogonais de Legendre.

A simulação dos dados foi imposta seguindo uma estrutura onde tanto a variância do efeito de ambiente permanente, como a variância do efeito de ambiente temporário, não fossem constantes ao longo do tempo. Para o efeito de ambiente temporário, foram consideradas duas variâncias diferentes, onde nas três primeiras idades mensuradas ela foi considerada igual a 2,15 unidades^2, e nas duas últimas foi admitida como sendo igual a 4,38 unidades^2.

Para a análise dos dados foram considerados dois diferentes modelos de regressão aleatória. No primeiro modelo (Modelo a) foram ajustadas funções de covariância tanto para o efeito genético aditivo, quanto para o efeito de ambiente permanente, onde ambas as funções utilizaram os três primeiros polinômios de Legendre, caracterizando uma função polinomial de segundo grau. Este modelo pode ser descrito por

$$\mathbf{y}_{ij} = \mathbf{F}_{ij} + \sum_{m=0}^{2} b_m f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} a_{im} f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} g_{im} f_m(\mathbf{a}^*_{ij}) + e_{ij} \quad \text{(Modelo a)}$$

em que $\mathbf{y}_{ij}$ é a **j**-ésima produção do **i**-ésimo indivíduo; $\mathbf{a}^*_{ij}$ é a idade na produção padronizada entre –1 a +1; f_m é o **m**-ésimo polinômio de Legendre; $\mathbf{F}_{ij}$ é o efeito fixo de local; b_m são os coeficientes de regressão para modelar a trajetória média comum a todos os indivíduos; a_{im} e g_{im} são os coeficientes de regressão aleatória dos efeitos genético aditivo e de ambiente permanente do indivíduo **i**, respectivamente, e e_{ij} é o efeito do ambiente temporário.

Em notação matricial o Modelo a é representado como:

$$\mathbf{y} = \mathbf{X}b + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \mathbf{e}$$

em que **y** é um vetor com as **n** observações de produção em cada idade; **X** é a matriz de incidência de níveis do efeito fixo de local e das idades padronizadas entre –1 a +1, associadas aos polinômios de Legendre, que descrevem a trajetória média de todos os indivíduos; **ß** é o vetor de soluções para níveis do efeito fixo de local e das soluções da regressão média de todos indivíduos; **Z** e **W** são matrizes de covariáveis referentes às idades padronizadas, associadas aos polinômios de Legendre, em cada produção, referentes aos coeficientes de regressão aleatória dos efeitos aleatórios genético aditivo e de ambiente permanente para cada indivíduo, respectivamente; **a** e **p** são vetores contendo os coeficientes de regressão aleatória para cada indivíduo, para os efeitos genético aditivo e de ambiente permanente, respectivamente; o vetor **e** representa os efeitos aleatórios de ambiente temporário. As pressuposições da distribuição dos vetores **a**, **p** e **e**, são as seguintes:

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} \sim \mathbf{N(0,V)} \text{ com } \mathbf{V} = \begin{bmatrix}\mathbf{A} \otimes \mathbf{Ka} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I} \otimes \mathbf{Kp} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{bmatrix},$$

sendo **A,** a matriz de coeficientes de parentesco entre indivíduos, de ordem igual ao número de indivíduos (**N**); **Ka** é a matriz de covariância entre os coeficientes de regressão aleatória do efeito genético aditivo; **Kp** é a matriz de covariância entre os coeficientes de regressão aleatória que descrevem o efeito de ambiente permanente; **I** é uma matriz identidade, de ordem igual ao número total de observações (**n**) e **R** é uma matriz diagonal de variância residual associada a cada observação.

O segundo modelo de regressão aleatória (Modelo b) diferiu do Modelo a por considerar o efeito de ambiente permanente constante, o que equivale a pressupor que o efeito de ambiente permanente é o mesmo para todas as idades, podendo ser descrito por um intercepto. O Modelo b é descrito por

$$\mathbf{y_{ij}} = \mathbf{F_{ij}} + \sum_{\mathbf{m=0}}^{2} b_{\mathbf{m}} f_{\mathbf{m}}(\mathbf{a}^{*}_{\mathbf{ij}}) + \sum_{\mathbf{m=0}}^{2} a_{\mathbf{im}} f_{\mathbf{m}}(\mathbf{a}^{*}_{\mathbf{ij}}) + g_{\mathbf{im}} + e_{\mathbf{ij}} \text{ (Modelo b),}$$

em que $g_{\mathbf{im}}$ representa o efeito do ambiente permanente. Os demais termos do modelo são definidos como no Modelo a. Para o Modelo b tem-se que

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} \sim \mathbf{N(0,V)} \text{ com } \mathbf{V} = \begin{bmatrix}\mathbf{A} \otimes \mathbf{Ka} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{bmatrix},$$

em que **P** é uma matriz diagonal de variância do efeito de ambiente permanente ($s^{2}_{\mathbf{P}}$), comum a todos os indivíduos.

Os dados simulados foram analisados adotando-se os Modelos a e b, sob duas situações. Na primeira situação considerou-se que o efeito de ambiente temporário apresentou duas variâncias diferentes ao longo do tempo, ou seja, nas três primeiras idades esse efeito apresentou menor variância que nas duas últimas idades. Na segunda

situação a variância do efeito de ambiente temporário foi assumida constante ao longo do tempo. Sendo assim, os Modelos a e b, combinados com as duas pressuposições diferentes com relação ao efeito de ambiente temporário originaram quatro modelos diferentes. No Modelo 1 ajustou-se uma função de covariância tanto para o efeito genético aditivo, quanto para o efeito de ambiente permanente, assumindo a variância do efeito de ambiente temporário heterogênea. O Modelo 2 foi similar ao Modelo 1, exceto que a variância do efeito de ambiente temporário foi considerada constante. Para os Modelos 3 e 4 o efeito de ambiente permanente foi assumido constante em toda trajetória, sendo que no Modelo 3 a variância do efeito de ambiente temporário foi assumida heterogênea e no Modelo 4 foi considerada homogênea. De posse dos resultados das análises, foi possível avaliar a importância tanto do efeito de ambiente permanente, como do efeito de ambiente temporário, quando são considerados constantes ou heterogêneos ao longo das idades avaliadas.

Os quatro modelos foram comparados por meio do teste da razão de verossimilhança e por meio do critério de informação de Akaike (AIC). O teste da razão de verossimilhança (Rao, 1973) permite comparar dois modelos hierárquicos. Assim, para a comparação do modelo **i**, que contém **n** parâmetros aleatórios, com o modelo **j**, que é o modelo reduzido com **p** parâmetros, a estatística do teste é dada por $\mathbf{LR_{ij} = 2Log_e L_i - 2Log_e L_j}$, em que $\mathbf{LR_{ij}}$ representa a estatística do teste de razão de verossimilhança para modelos seqüencialmente reduzidos; $\mathbf{Log_eL_i}$ e $\mathbf{Log_eL_j}$ são o logarítmo natural da função de verossimilhança restrita dos modelos **i** e **j**, respectivamente. A estatística de teste apresenta assintoticamente distribuição qui-quadrado, com $(\mathbf{n - p})$ graus de liberdade. A hipótese de nulidade testada consiste na igualdade dos valores das funções de verossimilhança restrita entre os modelos comparados, H_0: $\mathbf{2Log_eL_i=2Log_eL_j}$, ou seja, os parâmetros adicionais não reduzem a variância residual ou não são necessários. O critério de Akaike é expresso por: $\mathbf{AIC = -2\log_e L + 2p}$, em que **L** é o logarítmo da função de máxima verossimilhança restrita e **p** é o número de parâmetros estimados para o modelo em questão. Este critério

foi proposto por Akaike (1974) justamente para selecionar, dentro de um conjunto de modelos, aquele com o melhor ajuste. O importante neste critério é a diferença entre os seus valores, e não o tamanho dessa diferença. O melhor modelo é aquele que apresenta o menor valor de AIC.

Todas as análises foram realizadas no aplicativo DFREML Versão 3.0 α (MEYER, 1998) e no aplicativo Statistical Analysis System (SAS, 1990).

RESULTADOS E DISCUSSÃO

Descrição dos dados

A descrição dos dados, com as médias e desvios-padrão para a variável produção, por local e por idade, é apresentada na Tabela 1.

Tabela 1: Médias ($\overline{\mathbf{X}}$) e desvios padrão (DP) para a produção em cada idade e local

	Local 1	Local 2	Local 3	Total
Idades	$\overline{X}$ ± DP	$\overline{X}$ ± DP	$\overline{X}$ ± DP	$\overline{X}$ ± DP
12	14,97 ± 2,02	15,32 ± 1,93	15,16 ± 1,97	15,15 ± 1,98
30	26,78 ± 2,07	26,86 ± 1,98	26,70 ± 1,86	26,78 ± 1,97
48	35,79 ± 2,20	35,67 ± 2,24	35,80 ± 2,23	35,75 ± 2,22
66	41,71 ± 2,70	41,87 ± 2,60	41,76 ± 2,77	41,78 ± 2,69
84	45,01 ± 3,21	45,44 ± 3,24	45,24 ± 3,03	45,23 ± 3,16
Total	32,85 ± 11,16	33,04 ± 11,15	32,94 ± 11,15	32,94 ± 11,15

Neste estudo o interesse principal está nos efeitos aleatórios do modelo, e por isso foram simulados dados de tal forma que o efeito fixo de local fosse não significativo. Como apenas as proles possuem a informação da produção, cada uma das médias apresentadas no interior da Tabela 1 refere-se à média de 300 indivíduos mensurados em 5 idades, em cada local. Portanto, para a análise foram computados 120 progenitores e 900 indivíduos avaliados em 5 idades diferentes, totalizando 4500 mensurações.

Comparação de modelos

Na Tabela 2 estão apresentados os valores do logaritmo da função de máxima verossimilhança restrita (LogL), o resultado do Teste da Razão de Verossimilhança, o valor do Critério de Akaike (AIC) e os valores estimados para a(s) variância(s) residual(ais) para os quatro modelos estudados. Verificou-se que, ao analisar um conjunto de dados que apresenta variâncias de efeitos temporários heterogêneas (Modelos 1 e 3),

assumindo a pressuposição de que as mesmas são homogêneas (Modelos 2 e 4), as variâncias do efeito temporário estimadas para os Modelos 2 e 4 se aproximaram da média ponderada pelo número de observações entre as duas variâncias estimadas pelos Modelos 1 e 3. Resultado semelhante foi encontrado por OLORI et al. (1999), em avaliação genética de gado de leite. Então, se a amplitude entre as variâncias heterogêneas é muito alta, o efeito da violação dessa pressuposição (heterogeneidade de variâncias do efeito de ambiente temporário) sobre o ajuste do modelo empregado poderá ser maior. A implicação de considerar erroneamente a homogeneidade de variâncias, tanto do efeito de ambiente permanente, quanto do ambiente temporário, está na falta de ajuste do modelo empregado, que conseqüentemente faz com que a função de máxima verossimilhança estimada não seja aquela que minimiza o erro, comprometendo as estimativas e predições do modelo linear misto.

Tabela 2: Valores da estatística LogL, do Teste da Razão de Verossimilhança, resultados do Critério de Akaike (AIC) e os valores estimados para a(s) variância(s) residual(ais) nos quatro modelos estudados.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
LogL	-5759,18$^{(a)*}$	-5813,12$^{(b)}$	-5804,35$^{(c)}$	-5855,88$^{(d)}$
p	14	13	9	8
AIC	11546,37	11652,23	11626,69	11727,75
$\hat{s}^2_{e1}$	2,16	3,01	2,41	3,23
$\hat{s}^2_{e2}$	4,40		4,78	

*Valores de LogL seguidos de letras diferentes diferem entre si pelo Teste da Razão de Verossimilhança ao nível de 5% de probabilidade

Considerando o Critério de Akaike (AIC) e o teste da razão de verossimilhança, o modelo mais adequado ao conjunto de dados foi o Modelo 1, seguido pelos Modelos 3, 2 e 4, em ordem de melhor ajuste. No Modelo 1, tanto a variância do efeito de ambiente permanente, como a variância do efeito de ambiente temporário foram consideradas heterogêneas, ou seja, mudam com a idade. De fato, este resultado foi coerente, pois, como comentado anteriormente, os dados foram gerados de modo que as variâncias do

efeito de ambiente permanente e temporário não fossem constantes ao longo das mensurações. Na área de melhoramento genético animal, por exemplo, existem estudos de avaliação genética em gado de leite que afirmam que a variância do efeito de ambiente temporário geralmente é heterogênea durante a lactação. Resultados assim foram encontrados por JENSEN (2001).

EL FARO (2002), em estudos de produção de leite no dia do controle para vacas da raça Caracu, relatou que o modelo que considerou homogeneidade de variâncias residuais mostrou-se inadequado.

ARAÚJO (2003) trabalhando com avaliação genética da produção de leite na raça Holandesa, por meio de modelos de regressão aleatória, verificou que o efeito de ambiente permanente não foi constante ao longo da trajetória da lactação.

Os resultados encontrados na literatura estão de acordo com os encontrados neste estudo e evidenciaram a importância de se considerar a heterogeneidade de variâncias de um efeito quando o mesmo existe. Além disso, verificou-se que os modelos mais adequados são aqueles onde se considerou a heterogeneidade de variâncias do efeito de ambiente temporário. O segundo modelo mais adequado indicado pelos testes foi aquele onde o efeito de ambiente temporário foi considerado heterogêneo, mas o efeito de ambiente permanente foi considerado homogêneo. Esse fato indica que, para os testes utilizados, considerar a heterogeneidade de variâncias do efeito de ambiente temporário foi mais importante do que a do efeito de ambiente permanente. Para o melhoramento vegetal, pela inexistência de informações, é importante o conhecimento do padrão de comportamento da variância, tanto do efeito de ambiente permanente como a do efeito temporário, para a característica estudada, a fim de que se faça a adoção do modelo mais adequado.

Pelo fato de o Modelo 1 ter sido indicado pelos testes como o modelo mais adequado, as comparações feitas a respeito da adoção de diferentes modelos serão analisadas com base nos resultados obtidos pelo Modelo 1.

Coeficientes de Regressão Aleatória

Na Tabela 3 encontram-se as estimativas das covariâncias entre os coeficientes de regressão aleatória (na diagonal e abaixo da diagonal) e as estimativas de correlações genéticas (acima da diagonal) entre os coeficientes de regressão aleatória para o efeito genético aditivo e de ambiente permanente, quando ajustada uma função, para os modelos estudados.

Tabela 3: Estimativas de (co)variâncias entre os coeficientes de regressão aleatória (na diagonal e abaixo da diagonal) e estimativas de correlações genéticas entre os coeficientes de regressão aleatória (acima da diagonal) para os efeitos genético aditivo e de ambiente permanente, quando ajustada uma função, para os modelos 1, 2, 3 e 4.

Modelo			a_0	a_1	a_2
Modelo 1	Ka	a_0	2,1713	0,6820	-0,1330
		a_1	0,5736	0,3254	0,6270
		a_2	-0,0073	0,0133	0,0013
			p_0	p_1	p_2
	Kp	p_0	1,7095	0,2050	-0,4780
		p_1	0,2109	0,6163	0,2100
		p_2	-0,4383	0,1155	0,4912
Modelo 2			a_0	a_1	a_2
	Ka	a_0	2,2411	0,6730	-0,0030
		a_1	0,5782	0,3295	0,7380
		a_2	-0,0002	0,0208	0,0024
			p_0	p_1	p_2
	Kp	p_0	1,5704	0,4200	-0,3950
		p_1	0,4364	0,6888	0,4440
		p_2	-0,3371	0,2508	0,4639
Modelo 3			a_0	a_1	a_2
	Ka	a_0	3,1303	0,5080	-0,4600
		a_1	0,9026	1,0085	0,2120
		a_2	-0,6007	0,1573	0,5438
Modelo 4			a_0	a_1	a_2
	Ka	a_0	3,9423	0,5700	-0,2520
		a_1	1,2859	1,2905	0,5250
		a_2	-0,3661	0,4365	0,5358

Embora os resultados dos testes tenham indicado como segundo melhor modelo, o Modelo 3, comparando-se os Modelos 1 (aquele que apresentou o melhor ajuste) e 2, a magnitude das variâncias e covariâncias entre os coeficientes de regressão aleatória para o efeito genético aditivo foi bastante próxima. Entretanto, em relação ao Modelo 1, esses mesmos componentes foram superestimados nos Modelos 3 e 4. Nesses dois modelos, as variâncias dos coeficientes de regressão aleatória também apresentaram-se bastante próximas entre si, mas as covariâncias foram diferentes. COBUCCI (2002), em estudos de avaliação genética em gado de leite, comparando os Modelos 2 e 4, também verificou a superestimação das variâncias dos coeficientes de regressão aleatória para o efeito genético aditivo, quando considerou que as variâncias ambientais (efeito de ambiente temporário e permanente) eram homogêneas durante as mensurações (Modelo 4).

Para o efeito de ambiente permanente nos Modelos 1 e 2, as variâncias foram bastante semelhantes, ao passo que as covariâncias e correlações diferiram entre esses dois modelos.

Alterações nas correlações genéticas entre os coeficientes de regressão aleatória também foram verificadas quando se fez a adoção de diferentes modelos. Entretanto, REKAYA et al. (1999), em avaliação genética de gado de leite, não verificaram alterações nas correlações entre os coeficientes de regressão aleatória para o efeito genético aditivo, ao incluir coeficientes de regressão aleatória para o efeito de ambiente permanente. Confrontando os Modelos 1 e 2, verificou-se que as correlações a_0a_1 foram muito próximas entre si. Porém, para o Modelo 2, a magnitude da correlação a_0a_2 foi bem menor, e a correlação a_1a_2 foi superestimada. Analisando os Modelos 3 e 4, com relação ao Modelo 1, houve subestimação das correlações a_0a_1 e a_1a_2, mas as correlações a_0a_2 foram superestimadas nesses dois modelos. De acordo com REKAYA et al. (1999), a utilização dos modelos de regressão aleatória também permite inferir sob aspectos genéticos dentro de uma trajetória de crescimento. Entretanto, a seleção com base nos componentes relacionados com diferentes fases da trajetória de crescimento é complexa, pois a associação entre esses componentes ainda não é bem conhecida.

Neste trabalho, ao considerar a variância do efeito ambiental constante, ocorreu uma distorção das correlações entre os coeficientes de regressão aleatória. Assim, a seleção com base em um desses parâmetros poderia apresentar uma resposta irreal sobre a variação nos demais parâmetros. Comumente a seleção dos indivíduos não é feita com base nos valores dos coeficientes de regressão aleatória, mas sim em funções dos coeficientes de regressão aleatória que fornecem o valor genético aditivo dos indivíduos em qualquer ponto da trajetória de crescimento.

Componentes da variância

O comportamento das estimativas da variância genética aditiva e do ambiente permanente está representado nas Figuras 1 e 2, respectivamente. A variância genética aditiva foi superestimada quando se adotou os Modelos 3 e 4, e nos Modelos 1 e 2 ela apresentou praticamente o mesmo comportamento dos dados simulados. Nos Modelos 3 e 4 a superestimação da variância do efeito genético aditivo ocorreu de forma acentuada, sendo crescente com o avanço da idade, exceto aos 12 meses (início da trajetória), onde a variância do efeito genético aditivo do Modelo 4 foi semelhante à do Modelo 1. O Modelo 4 apresentou superestimação ainda maior nas idades mais avançadas. Portanto, verificou-se que modelos que consideraram a variância do efeito de ambiente permanente constante foram os menos indicados. Entretanto, de acordo com o critério de Akaike, o segundo melhor modelo foi o Modelo 3, onde a variância do efeito de ambiente temporário foi heterogênea, mas a variância do efeito de ambiente permanente foi homogênea. Retornando ao estudo do comportamento da variância do efeito genético aditivo, a pior situação ocorreu quando, além da variância do efeito de ambiente permanente, a variância do efeito de ambiente temporário também foi considerada constante (Modelo 4). Esses resultados foram concordantes com alguns estudos em avaliação genética de gado de leite, como o de REKAYA et al. (1999), KETTUNEN et al. (2000) e COBUCCI (2002), que também verificaram que a variância do efeito genético é sempre superestimada quando se considera que as variâncias do efeito de ambiente

permanente e temporário são homogêneas ao longo das idades. Porém, neste trabalho verificou-se que, para a estimativa da variância genética aditiva, considerando-se a heterogeneidade da variância do efeito de ambiente permanente, considerar ou não a heterogeneidade do efeito de ambiente temporário não promoveu diferenças. Porém, quando não se considerou a heterogeneidade do efeito de ambiente permanente, deixar de considerar também a variância do efeito de ambiente temporário resultou em superestimação ainda maior da variância genética aditiva nas idades mais avançadas. De acordo com COBUCCI (2002), são necessários estudos adicionais para verificar o efeito de se considerar a variância do efeito de ambiente temporário heterogênea, em estudos de curvas de lactação, o que poderia levar a uma estimação mais acurada dos parâmetros genéticos.

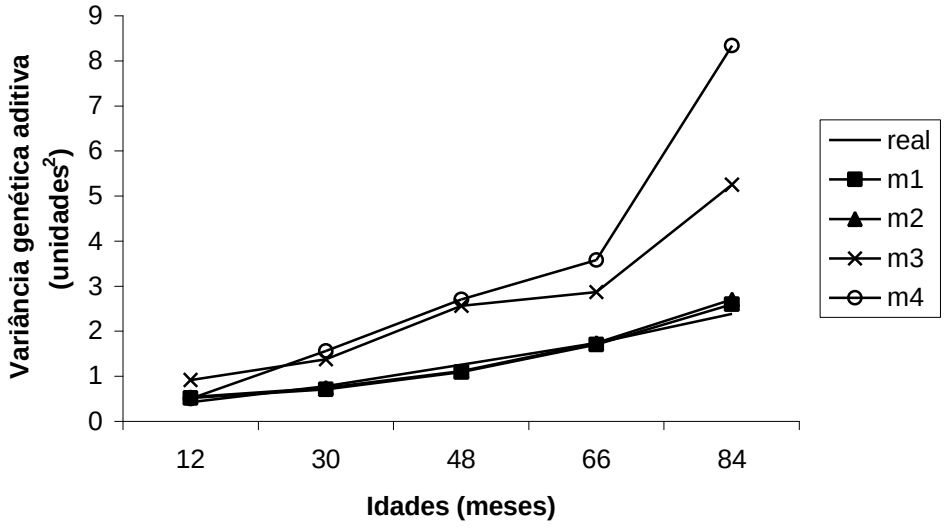

Figura 1 – Representação gráfica do comportamento da estimativa da variância genética aditiva nos modelos 1, 2, 3 e 4 e para os dados simulados em cinco idades de avaliação.

EL FARO e ALBUQUERQUE (2003) utilizaram diferentes modelos de regressão aleatória, comparando o uso de classes de variâncias com uso de funções de variância para descrever a variância do efeito de ambiente temporário. Modelos com classes de

variâncias residuais foram superiores àqueles com função de variância, enquanto que o modelo assumindo a homogeneidade de variância foi inadequado. BERMEJO et al. (2003) procurando descrever a variação genética do consumo de alimentos em suínos, por meio de modelos de regressão aleatória, verificaram que combinar uma estrutura de matriz do tipo diagonal para descrever o efeito de ambiente permanente, associada à estrutura do tipo auto-regressiva para descrever a variância do efeito de ambiente temporário, produziu o modelo de regressão aleatória mais adequado.

Como comentado anteriormente, os Modelos 1 e 2 foram semelhantes na estimação da variância genética aditiva. Entretanto, esses modelos diferiram quanto à estimação do efeito de ambiente permanente. As estimativas do Modelo 1 foram semelhantes às dos dados simulados sendo levemente superestimada à 30, 48 e 66 semanas. Já as estimativas do Modelo 2 foram subestimadas no início (à 12 semanas) e superestimadas no final da trajetória (à 66 e 84 semanas). Foi evidente que, quando se considerou a variância do efeito de ambiente permanente constante (Modelos 3 e 4), a estimativa desse componente foi subestimada em toda a trajetória, e quando se considerou a variância do efeito de ambiente temporário também constante, a subestimação da variância foi ainda menor.

Estes resultados indicaram que, como em programas de melhoramento genético o componente de maior importância é o componente genético aditivo, o modelo que considerou apenas a heterogeneidade da variância do efeito de ambiente permanente descreveu a variância do efeito genético aditivo tão bem quanto o modelo que considerou conjuntamente as variâncias do efeito de ambiente permanente e temporário heterogêneas.

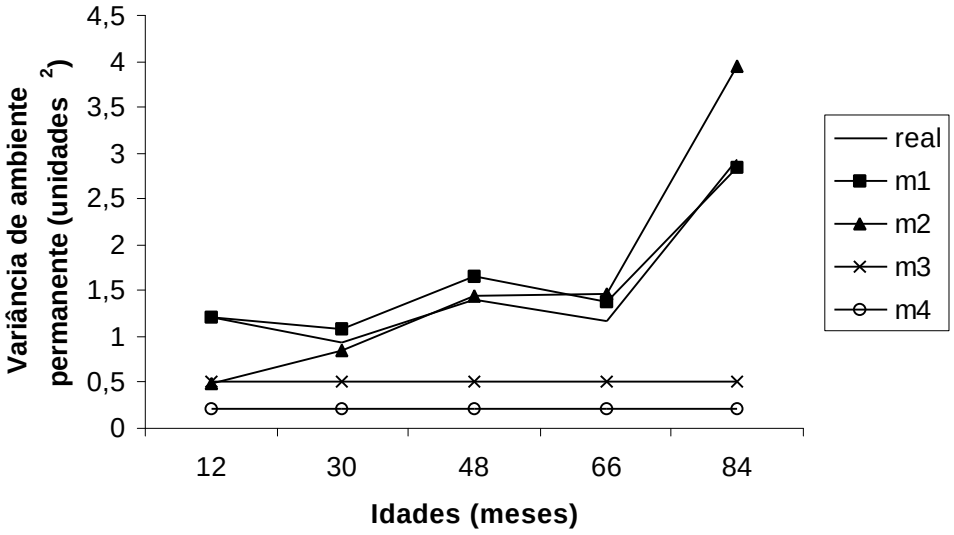

Figura 2 – Representação gráfica do comportamento da estimativa da variância do efeito de ambiente permanente nos modelos 1, 2, 3 e 4 e para os dados simulados em cinco idades de avaliação.

Parâmetros genéticos

Os dados foram gerados de modo a apresentar variâncias genéticas crescentes ao longo das mensurações. Já a variância ambiental (que compreende as variâncias do efeito de ambiente permanente e temporário) foi programada para ser menor nas idades iniciais e maior nas idades mais avançadas. A representação gráfica das estimativas de herdabilidade ao longo das mensurações está apresentada na Figura 3. Os valores de herdabilidade variaram de 0,13 a 0,26 no Modelo 1; 0,13 a 0,28 no Modelo 2; 0,24 a 0,50 no Modelo 3 e 0,13 a 0,71 no Modelo 4. Dado que os valores reais de herdabilidade variaram de 0,11 a 0,26, de modo geral, os Modelos 3 e 4 superestimaram a herdabilidade, principalmente em idades mais avançadas. A trajetória dos valores reais de herdabilidade foi crescente nas 3 primeiras idades e a partir daí tornou-se constante.

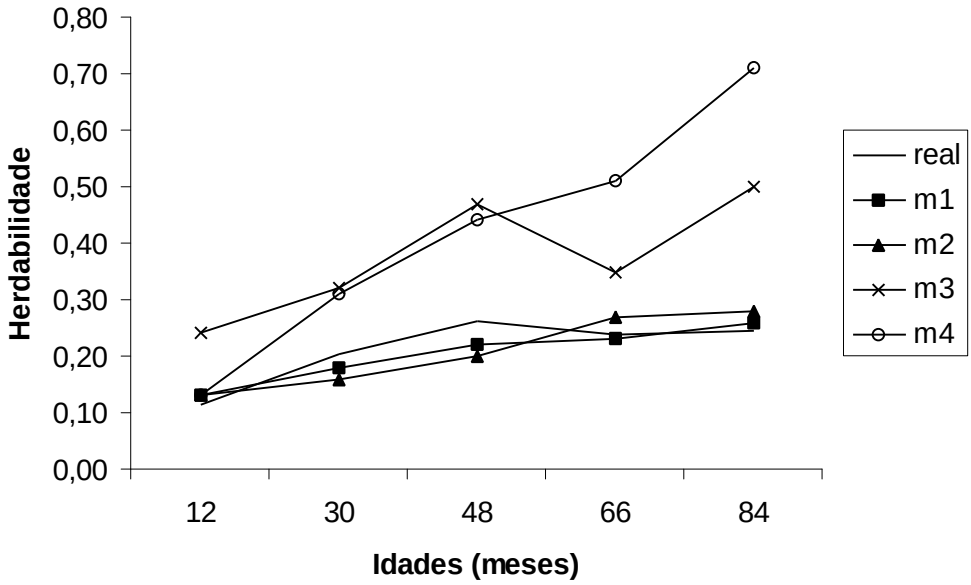

Figura 3 – Representação gráfica do comportamento da estimativa da herdabilidade estimada pelos Modelos 1, 2, 3 e 4 e para os dados simulados em cinco idades de avaliação.

O comportamento das estimativas de herdabilidade ao longo da trajetória foi semelhante ao comportamento verificado para a variância do efeito genético aditivo. Os verdadeiros valores de herdabilidade e as estimativas obtidas por meio dos Modelos 1 e 2 foram semelhantes. Os pequenos desvios apresentados entre os Modelos 1 e 2 e os verdadeiros valores provavelmente se devem às flutuações dos componentes da variância já citadas anteriormente. Por exemplo, as estimativas de herdabilidade obtidas no Modelo 1 foram subestimadas em relação aos verdadeiros valores nas idades 30, 48 e 66 semanas, que são exatamente os pontos em que a variância do efeito de ambiente permanente foi superestimada.

REKAYA et al. (1999), comparando os Modelos 3 e 4 em avaliação genética de gado de leite, verificaram que as estimativas de herdabilidade foram sempre maiores para o Modelo 4. De acordo com os autores, isto provavelmente se deve à partição incorreta dos componentes genéticos e ambientais, causada pela homogeneidade de variâncias

ambientais, assumida incorretamente. Em outras palavras, por meio do modelo adotado, não foi possível distinguir corretamente a parte da variância genética da parte da variância ambiental.

Uma conseqüência prática disto, é que, ao selecionar 20% dos melhores indivíduos, resultando em uma intensidade de seleção de 1,40, os ganhos genéticos esperados aos 12 e aos 84 meses na situação real seriam de 0,3048 e 1,1056 unidades, respectivamente. No Modelo 1 os ganhos esperados nas mesmas idades seriam de 0,3603 e 1,1499 unidades, respectivamente, enquanto que para o Modelo 4 os ganhos esperados seriam de 0,3603 e 3,0959, respectivamente. Portanto, ignorar alterações da covariância ambiental ao longo da trajetória pode resultar em superestimação da estimativa da herdabilidade. Assim, quando a variação ambiental é modelada corretamente por meio de uma estrutura de variâncias e covariâncias mais adequada, obtém-se estimativas de parâmetros genéticos mais acuradas.

O padrão de correlação genética da produção entre as idades demonstrou que idades mais próximas são mais correlacionadas entre si. As correlações dos dados gerados foram todas positivas e superiores a 0,50 (Figura 4a). Quando se ajustou uma função de covariância para os efeitos genético aditivo e de ambiente permanente (Modelos 1 e 2), a função de covariância que descreveu a variação genética suavizou as estimativas de covariâncias, acompanhando de forma bem similar o padrão verdadeiro (Figuras 4b e 4c). Admitir o efeito de ambiente permanente constante em toda trajetória (Modelo 3), superestimou as variâncias genéticas em cada idade e subestimou as covariâncias genéticas entre idades mais distantes, ocorrendo inclusive valores negativos, resultando assim, em correlações genéticas subestimadas (Figura 4d). Considerar tanto a variância do efeito de ambiente permanente, quanto a do ambiente temporário, constantes (Modelo 4), superestimou ainda mais a variância genética aditiva naquelas idades, onde o efeito do ambiente permanente é maior. As covariâncias entre idades mais distantes da trajetória ficaram menos subestimadas do que no Modelo 3, porém a superestimação das variâncias foi maior, ocasionando correlações genéticas menos subestimadas (Figura 4e), quando comparadas àquelas obtidas pelo Modelo 3.

KETTUNEN et al. (2000) observaram menores valores de correlação genética entre controles de produção de leite, principalmente em pontos mais distantes, quando compararam o modelo de regressão aleatória que considerou constante o efeito de ambiente permanente, com outro modelo que ajustou uma função de covariância para este efeito. Padrão similar também foi verificado por REKAYA et al. (1999) e COBUCI (2002).

As características que se repetem no mesmo indivíduo possuem intensidades diferentes a cada ocasião em que se expressam, devendo-se esta variação a fatores de ambiente temporário, visto que as medidas são feitas no mesmo genótipo. Já variações entre indivíduos são decorrentes tanto de fatores genéticos, como ambientais, neste caso, de ambiente permanente. Ao considerar o efeito de ambiente permanente constante, quando na verdade ele é heterogêneo, resulta em falta de ajuste na descrição desse efeito ao longo da trajetória, o que pode ser erroneamente contabilizado como diferenças entre indivíduos provenientes da ação genética aditiva, superestimando a variância genética aditiva e, conseqüentemente, as estimativas de herdabilidade.

Considerar a estrutura de (co)variâncias dos efeitos de ambiente permanente e ambiente temporário melhora a capacidade do modelo em distinguir com maior precisão a variação causada no fenótipo dos indivíduos, proveniente dos efeitos de origem genética, daqueles de origem não genética.

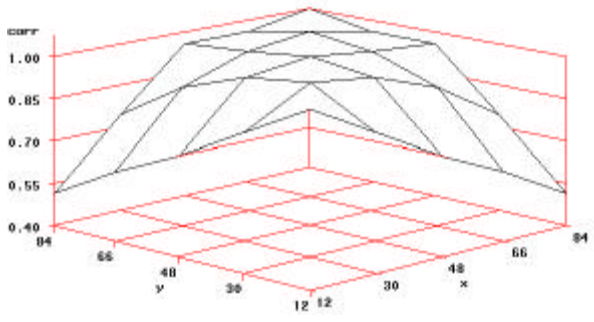

Figura 4 a

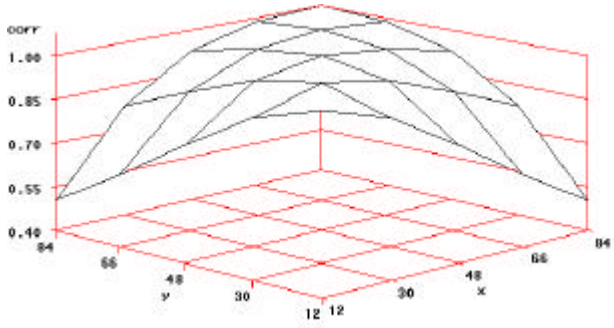

Figura 4 b

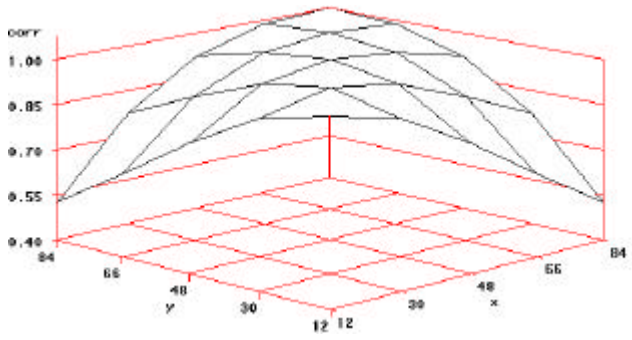

Figura 4 c

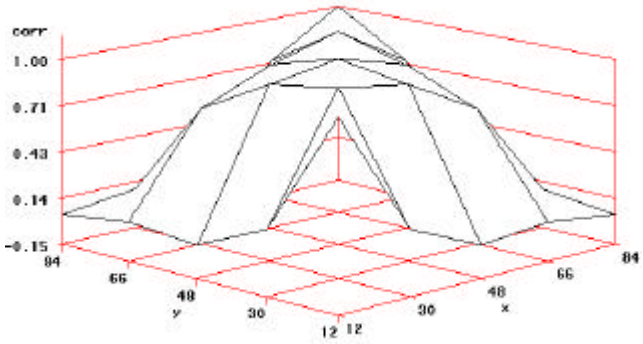

Figura 4 d

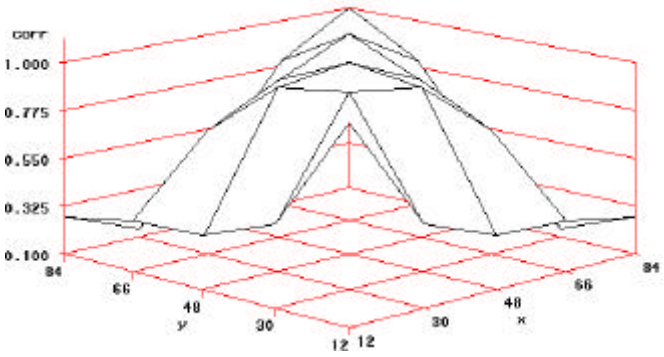

Figura 4 e

Figura 4 – Correlações genéticas entre as idades para os dados simulados (4 a) e suas estimativas obtidas pelos Modelos 1 (4 b), 2 (4 c), 3 (4 d) e 4 (4 e).

Predição dos valores genéticos

Por meio das soluções da regressão aleatória obtidas para o efeito genético aditivo, foram preditos os valores genéticos dos indivíduos em cada idade, para cada modelo. Para cada idade foi obtida a correlação de ordem dos valores genéticos preditos pelos modelos, para uma amostra de 10% dos melhores indivíduos e também para todos os indivíduos (Tabela 4). Verificou-se que a correlação obtida quando todos os indivíduos foram considerados foi sempre maior que 0,78. A porcentagem de coincidência entre os Modelos 1 e 2 sempre se aproximou de 100% em todas as idades, independente do tamanho da amostra. Porém, considerando 10% dos melhores indivíduos, a porcentagem de coincidência entre o Modelo 1 e o Modelo 3 variou de 0,25 (aos 12 meses) a 0,69 (aos 66 meses), e entre o Modelo 1 e o Modelo 4 variou de 0,40 (aos 84 meses) a 0,70 (aos 66 meses), indicando que os melhores indivíduos selecionados pelo Modelo 1, teriam sua classificação alterada nos Modelos 3 e 4.

Ao confrontar todos os resultados, verificou-se pelos valores obtidos por meio do teste da razão de verossimilhança e o critério de informação de Akaike, que assumir o efeito de ambiente temporário como heterogêneo foi mais importante do que considerar a mesma pressuposição para o efeito de ambiente permanente. Isto se deve à maior proporção da variância do efeito temporário sobre a variância total (fenotípica), quando comparada à parcela da variância do efeito de ambiente permanente. Conseqüentemente, os testes apontaram o Modelo 3 com melhor qualidade de ajuste do que o Modelo 2. Por outro lado, ao confrontar ambos os modelos na estimação da variância genética aditiva, herdabilidade, na obtenção da correlação genética entre as idades e no percentual de coincidência em selecionar os melhores indivíduos com base em seus méritos genéticos, o Modelo 2 foi o que mais se aproximou dos resultados obtidos pelo Modelo 1, e conseqüentemente, ao padrão verdadeiro dos dados. Assim, enquanto o Modelo 3

proporcionou maior precisão, o Modelo 2, proporcionou maior acurácia. Sendo assim, o Modelo 1, além de fornecer maior precisão, também forneceu maior acurácia.

Tabela 4: Correlação de Spearmam (de ordem) entre os valores genéticos preditos para uma amostra de 10% dos melhores indivíduos (acima da diagonal) e para todos os indivíduos da população (abaixo da diagonal)

12 meses	Mod1	Mod2	Mod3	Mod4
Mod1	1,00	0,98	0,25	0,59
Mod2	0,99	1,00	0,21	0,58
Mod3	0,79	0,78	1,00	0,76
Mod4	0,95	0,94	0,90	1,00

30 meses	Mod1	Mod2	Mod3	Mod4
Mod1	1,00	0,98	0,39	0,45
Mod2	0,99	1,00	0,35	0,45
Mod3	0,85	85,00	1,00	0,95
Mod4	0,85	0,89	0,98	1,00

48 meses	Mod1	Mod2	Mod3	Mod4
Mod1	1,00	0,98	0,53	0,66
Mod2	0,99	1,00	0,45	0,62
Mod3	0,83	0,83	1,00	0,91
Mod4	0,87	0,87	0,98	1,00

66 meses	Mod1	Mod2	Mod3	Mod4
Mod1	1,00	0,98	0,69	0,70
Mod2	0,99	1,00	0,63	0,72
Mod3	0,93	0,93	1,00	0,81
Mod4	0,93	0,94	0,97	1,00

84 meses	Mod1	Mod2	Mod3	Mod4
Mod1	1,00	0,97	0,45	0,40
Mod2	0,99	1,00	0,51	0,48
Mod3	0,89	0,89	1,00	0,97
Mod4	0,85	0,86	0,98	1,00

CONCLUSÕES

Modelos de regressão aleatória por permitirem ajustar a estrutura de covariância mais adequada aos dados amostrais, permitem obter parâmetros genéticos mais precisos, e conseqüentemente, predições de efeitos aleatórios mais acuradas.

No uso de modelos de regressão aleatória para descrição da variação de dados longitudinais, deve-se considerar a estrutura da variância residual, visto que assumir pressuposições incorretas sobre a estrutura de covariância dos efeitos aleatórios do modelo afeta a estrutura de covariância, comprometendo as estimativas de parâmetros genéticos. Ignorar a heterogeneidade de variâncias residuais resultou na falta de ajuste do modelo adotado, que conseqüentemente fez com que a função de máxima verossimilhança estimada não fosse aquela que minimiza o erro, comprometendo as estimativas e predições do modelo linear misto. Assim, a seleção dos melhores indivíduos poderia estar sendo realizada mais em função do ambiente do que devido a diferenças genéticas entre os indivíduos.

Ignorar a heterogeneidade de variâncias do efeito de ambiente permanente e analisar dados que apresentam variâncias de efeitos temporários heterogêneas, assumindo a pressuposição de que as mesmas são homogêneas, superestima as variâncias genéticas em cada ponto na trajetória e subestima as covariâncias genéticas entre pontos mais distantes.

REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO,C. V. Modelos de Regressão Aleatória para avaliação Genética da produção de leite na raça Holandesa. Viçosa. UFV, 2003, 85p. Tese (Doutorado em Zootecnia) – Universidade Federal de Viçosa, M.G., 2003.

AKAIKE, H. A new look at the statistical model identification. Trans. Automat. Contr. n. 19,p. 716-723, 1974.

BERMEJO, J. L.; ROECHE, R; SCHULZE, V.; et al. 2003. Random regression to model genetically teh longitudinal data of daily feed intake in growing pigs. *Livest. Prod. Sci*., 82:189-199.

COBUCCI, J.A. 2002. Uso de Modelos de Regressão Aleatória na Avaliação da Persistência na Lactação de animais da Raça Holandesa. Viçosa: UFV, 2002, 99p. Tese (Doutorado em Zootecnia) – Universidade Federal de Viçosa, M.G., 2002.

EL FARO, L. 2002. Estimação de componentes de (co)variância para a produção de leite no dia de controle de primeiras lactações de vacas Caracu, aplicando-se “test day models” de dimansão finita e modelos de regressão aleatória**.** Jaboticabal: UNESP, 102 p. Tese (Doutorado em Zootecnia) – Universidade Estadual Paulista, SP, 2002.

EL FARO, L.; ALBUQUERQUE, L. G. 2003. Utilização de modelos de regressão aleatória para produção de leite no dia de controle, com diferentes estruturas de variâncias residuais. *Rev. Bras. Zoot.*32(5): 345-352.

JAMROZIK, J., SCHAEFFER, L.R. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. *J. Dairy Sci.*, 80(4):762-770.

JAMROZIK, J., SCHAEFFER, L.R., DEKKERS, J.C.M. 1997. Genetic evaluation of dairy cattle using test day yields and random regression model. *J. Dairy Sci.*, 80(6):1217-1226.

JENSEN, J. Genetic evaluation of dairy cattle using test-day models. 2001. *J. Dairy Sci.*, 84(12):2803-2812.

KETTUNEN, A., MÄNTYSAARI, E.A., POSO, J. 2000. Estimation of genetic parameters for daily milk yield of primiparous Ayrshire cows by random regression test-day models. *Livest. Prod. Sci.*, 66:251-261.

MEYER, K. 2004. Scope for a random regression model in genetic evaluation of beef cattle for growth. *Livest. Prod. Sci.*, 86:69-83.

MEYER, K. DXMRR – A set programs to estimate COVARIANCE FUNCTIONS FOR LONGITUDINAL DATA BY REML. In: World Congress of Genetics Applied to Livestock Production, 6, 1998, Armidale. Proceeding… Armidale: University of New England, 1998. CD-ROM.

MEYER, K. 2000. Random regressions to model phenotypic variation in monthly weights of Australian beef cows. *Livest. Prod. Sci.*, 65:19-38.

OLORI, V.E., HILL, W.G., McGUIRK, B.J., BROTHERSTONE, S. 1999. Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livest. Prod. Sci.*, 61:53-63.

RAO, C.R. 1973. Linear statistical inference and its applications. 2ed. New York:John Wiley & Sons. 552p.

REKAYA, R., CARABAÑO, M.J., TORO, M.A. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. *Livest. Prod. Sci.*, 57:203-217.

SAS INSTITUTE INC. SAS/STAT® user's guide, version 6. 4ed. Carry, NC. 1990. v.1, 943p.

SCHAEFFER, L.R., DEKKERS, J. C. M. 1994 Random regression in animal models for test day production in dairy cattle. In: World congress genetic applied livestock production, 5., 1994, Guelph. ON, Canada, *Proccedings*… Guelph, 1994. p. 443-446.

Capítulo 4

Estimativas de componentes de variância e parâmetros genéticos em diferentes estratégias de análise de dados longitudinais obtidas por meio de simulação.

Resumo- Foi gerada uma estrutura de pedigree, na qual 30 progenitores masculinos foram cruzados com três progenitores femininos diferentes cada um, originando dez proles diferentes em cada cruzamento, distribuídas em três locais diferentes. Para cada prole foram geradas informações fenotípicas em cinco idades diferentes, ou seja, aos 12, 30, 48, 66 e 84 meses. O arquivo de dados foi gerado permitindo que algumas informações fossem perdidas, resultando em arquivos de dados com 10, 20, 30 e 40% de perda de informação. O mesmo procedimento foi realizado novamente, porém, os indivíduos com informações perdidas foram os de menor valor fenotípico, representando o efeito da seleção. Modelos unicaracterística, de repetibilidade, de regressão aleatória e multi-característica foram utilizados para a análise dos dados. Modelos de regressão aleatória foram mais adequados para descrever continuamente as estruturas de covariâncias de crescimento ao longo do tempo, do que modelos unicaracterística e de repetibilidade, quando a pressuposição de que a correlação entre mensurações sucessivas no mesmo indivíduo é diferente da unidade. Sob ausência de seleção, os modelos de regressão aleatória e multi-característica foram semelhantes. Entretanto, sob o efeito da seleção, o modelo multi-característica mostrou-se mais susceptível ao viés de seleção do que o modelo de regressão aleatória.

Palavras-chave: modelo de regressão aleatória, modelo multi-característica, modelos de repetibilidade, seleção.

Chapter 4

Estimates of variance components and genetic parameters in different analysis strategies in longitudinal data obtained by simulation

Abstract: A pedigree structure was generated where 30 male progenitors were crossed with 3 female progenitors each one, generating ten offspring in each cross, distributed in three different places. For each offspring, phenotypic information was generated in five different ages, at 12, 30, 48, 66 and 84 months. The data file was simulated allowing some information to be lost, resulting in data with 10, 20, 30 and 40% of lost information. The same procedure was performed, but the individuals with lost information, were the ones with lower phenotypic values, representing the selection effect. The single-trait model, repeatability model, random regression model and multiple-trait model were used to analyze the data. Random regression models were more adequate to describe continually the covariance structure of growth over time, than single-trait and repeatability models, when the assumption of correlation between successive measurements in the same individual is different from one. Without selection, random regression model and multiple-trait models were very similar. However, under selection effect, multiple-trait model showed more susceptible to bias than the random regression model.

Key words: random regression model, multiple-trait model, repeatability model, selection.

INTRODUÇÃO

Dados oriundos de sucessivas medições em um mesmo indivíduo (dados longitudinais) podem ser analisados por diferentes estratégias. No melhoramento genético uma abordagem que tem apresentado características bastante interessantes é a utilização de modelos de regressão aleatória. HENDERSON JUNIOR (1982) propôs a teoria a respeito dos chamados coeficientes de regressão aleatória, baseando-se no princípio de que, se existe um coeficiente de regressão pertencente a cada indivíduo em um experimento, e se há uma amostra aleatória de indivíduos, então os coeficientes de regressão devem ser considerados aleatórios. Esta metodologia tem sido utilizada para modelar características que são medidas no tempo, tais como produção de leite durante a lactação e características de crescimento. SCHAEFFER e DEKKERS (1994) propuseram a utilização dos coeficientes de regressão aleatória, no contexto do melhoramento genético, à dados de produção de leite no dia do controle. Isto permitiu que cada animal apresentasse uma curva de lactação individual, descrita por uma regressão fixa, comum a todos os animais, e regressões aleatórias para cada animal, descrevendo os desvios em relação à regressão fixa. Atualmente os modelos de regressão aleatória têm sido bastante utilizados para descrever curvas de lactação, que é a área onde a aplicação desses modelos está mais desenvolvida. Na adoção desse modelo, as medições ao longo do tempo são consideradas como pontos sucessivos sobre uma trajetória contínua, e é permitida a predição de parâmetros, inclusive em pontos (idades) onde não tenham sido realizadas mensurações. Para descrever tanto a curva fixa para todos os indivíduos, como as curvas individuais, funções de covariância (KIRKPATRICK et al., 1990), que descrevem a estrutura de covariância entre idades, podem ser utilizadas. Funções de covariância utilizando os polinômios de Legendre têm sido bastante utilizadas porque são os mais fáceis de se calcular e utilizar.

Outra alternativa de análise para dados de natureza longitudinal seria por meio dos modelos lineares simples. No modelo de repetibilidade, inclui-se o efeito de ambiente permanente, e assume-se a pressuposição de que as medidas sucessivas no mesmo

indivíduo apresentam correlação igual à unidade, ou seja, que as covariâncias entre duas medidas sucessivas são iguais. Entretanto, tal pressuposição nem sempre é válida, pois para características de crescimento, por exemplo, medidas mais próximas são, muitas vezes, mais fortemente correlacionadas entre si do que medidas mais distantes no tempo. Na abordagem de características múltiplas não é feita qualquer pressuposição a respeito das covariâncias (a matriz de covariâncias é do tipo não estruturada). Além disso, esse modelo é altamente parametrizado, pois requer a matriz completa de covariâncias e, quando se tem um grande número de idades avaliadas, o número de parâmetros a serem estimados torna-se muito grande, levando à dificuldades computacionais.

Existem na literatura, diversos trabalhos comparando os resultados quando se utilizam diferentes estratégias de análise para dados longitudinais (ARAÚJO, 2003; RESENDE et al., 2001, HUISMAN et al.,2001, KETTUNEN et al., 2000, REKAYA et al., 1999). Entretanto, ainda não foram verificados trabalhos considerando o comportamento desses modelos diante do desbalanceamento dos dados.

O objetivo deste trabalho foi verificar o comportamento das estimativas dos componentes de variância e dos parâmetros genéticos, quando são utilizados modelos de regressão aleatória, o modelo de repetibilidade e o modelo de multi-característica, quando se tem diferentes níveis de desbalanceamento dos dados. Para tal, foi realizada a análise desses três modelos com os dados completos e com 10, 20, 30 e 40% de desbalanceamento dos dados, respectivamente. O efeito da seleção sobre os dados também foi avaliado.

MATERIAL E MÉTODOS

Material

Foi gerada uma estrutura de pedigree onde 30 progenitores masculinos foram acasalados com três progenitores femininos diferentes cada um, originando dez proles diferentes em cada acasalamento, distribuídas em três locais diferentes. Para cada prole foram geradas informações fenotípicas em cinco idades diferentes, ou seja, aos 12, 30, 48, 66 e 84 meses. Assim, a geração dos dados resultou em 120 progenitores (30 masculinos e 90 femininos), 1020 indivíduos no total, sendo 900 indivíduos possuindo informação de produção em cinco idades diferentes, totalizando em 4500 registros de produção.

Os dados foram gerados para representar uma estrutura de dados longitudinais, de um conjunto de dados de **N** indivíduos, com cada indivíduo contendo registros em cinco idades diferentes (**i**=5). O modelo linear misto foi definido como: $\mathbf{y} = \mathbf{X}b + \mathbf{Za} + \mathbf{Wp} + e$, em que **y** é um vetor de observações dos indivíduos; b é um vetor de parâmetros de efeitos fixos e de parâmetros da curva geral para todos os indivíduos; **X** é uma matriz contendo a incidência dos níveis de efeitos fixos e as variáveis regressoras, compostas pelas idades padronizadas associadas aos Polinômios de Legendre; **a e p** são vetores aleatórios de soluções de regressão aleatória do efeito genético aditivo e de regressão aleatória do efeito de ambiente permanente, respectivamente; **Z** e **W** são matrizes que associam as variáveis regressoras aos vetores **a** e **p**, respectivamente. O vetor e é um vetor de efeito aleatório de ambiente temporário.

Assumindo que os vetores **y**, **a**, **p** e e apresentam distribuição normal, a esperança do vetor aleatório **a** é $\mathbf{E(a)} = \mathbf{0}$ e a variância é $\mathbf{V(a)} = \mathbf{A} \ddot{A} \mathbf{Ka} = \mathbf{G}$, onde **Ka** é uma matriz de covariâncias entre os coeficientes de regressão aleatória do efeito genético aditivo e **A** é a matriz que indica o grau de parentesco entre os indivíduos, de dimensão igual ao número total de indivíduos (**N**). O vetor **p** possui $\mathbf{E(p)} = \mathbf{0}$ e a variância é $\mathbf{V(p)} = \mathbf{I} \ddot{A} \mathbf{Kp} = \mathbf{P}$, onde **Kp** é uma matriz de covariâncias entre os coeficientes de

regressão aleatória do efeito de ambiente permanente e **I** uma matriz identidade de dimensão igual ao número de indivíduos com observações (**n**). Por último, e possui média $\mathbf{E(e) = 0}$ e variância $\mathbf{V(e) = I}\,s_e^2 = \mathbf{R}$, onde s_e^2 é a variância do efeito de ambiente temporário. Conseqüentemente, a esperança e a variância do vetor de dados observados **y**, são $\mathbf{E(Y) = X}b$ e $\mathbf{V(Y) = ZGZ' + WPW' + R}$, respectivamente.

Ao vetor b estão associados os efeitos fixos e a curva de regressão geral para todos os indivíduos. Se $\mathbf{UT_a^{1/2}}$ e $\mathbf{UT_p^{1/2}}$ são as decomposições de Cholesky das matrizes de covariâncias dos coeficientes de regressão aleatória dos efeitos genético aditivo e de ambiente permanente, respectivamente, e **A** é uma matriz de numeradores de coeficiente de parentesco entre os indivíduos, onde $\mathbf{AT_a^{1/2}}$ é a decomposição de Cholesky dessa matriz de parentesco, então o vetor **y** (fenótipos) contendo as "**i**" características (idades) é definido como:

$$\mathbf{y} = \mu + \mathbf{AT_a^{1/2}}\ \mathbf{Za}\ \ \mathbf{UT_a^{1/2}} + \mathbf{Zp}\ \ \mathbf{UT_p^{1/2}} + \mathbf{e}.$$

A simulação dos dados foi realizada por meio de um modelo polinomial de segundo grau, utilizando os polinômios ortogonais de Legendre para descrever tanto a trajetória fixa, quanto a dos efeitos aleatórios que compõem o modelo, representado como:

$$\mathbf{y_{ijk}} = \mathbf{L_j} + \mathbf{b_0} + \mathbf{b_1(A)} + \mathbf{b_2(A)^2} + (\mathbf{a_{i0}} + \mathbf{a_{i1}(A)} + \mathbf{a_{i2}(A)^2}) + (\mathbf{p_{i0}} + \mathbf{p_{i1}(A)} + \mathbf{p_{i2}(A)^2}) + \mathbf{e_{ijk}}$$

em que $\mathbf{L_j}$ é o efeito fixo de local, com **j**=1, 2 e 3; $\mathbf{b_o}$, $\mathbf{b_1}$ e $\mathbf{b_2}$ são coeficientes de regressão que descrevem a curva fixa geral da produção em função das idades padronizadas (**A**) entre –1 a +1 e associadas aos três primeiros polinômios de Legendre, $\mathbf{a_{i0}}$, $\mathbf{a_{i1}}$ e $\mathbf{a_{i2}}$ são os coeficientes de regressão aleatória relacionados ao efeito genético aditivo para o indivíduo **i**, assumidos tendo distribuição multinormal com vetor de média zero e matriz de covariâncias **Ka**; $\mathbf{p_{i0}}$, $\mathbf{p_{i1}}$ e $\mathbf{p_{i2}}$ são os coeficientes de regressão aleatória relacionados ao efeito de ambiente permanente para o indivíduo **i**, assumidos tendo

distribuição multinormal com vetor de média zero e matriz de covariâncias **Kp**; $\mathbf{e}_{ijk}$ é o efeito residual temporário assumido tendo distribuição normal com média zero e variância s_e^2=2,2 unidades^2.

As matrizes **Ka** e **Kp**, de covariâncias dos coeficientes de regressão aleatória do efeito genético aditivo e do efeito de ambiente permanente, respectivamente, utilizadas foram:

$$\mathbf{Ka} = \begin{bmatrix} \mathbf{2,2367} & \mathbf{0,5571} & \mathbf{-0,1227} \\ \mathbf{0,5571} & \mathbf{0,2939} & \mathbf{0,0027} \\ \mathbf{-0,1227} & \mathbf{0,0027} & \mathbf{0,0328} \end{bmatrix} \text{ e } \mathbf{Kp} = \begin{bmatrix} \mathbf{1,4140} & \mathbf{0,1405} & \mathbf{-0,3454} \\ \mathbf{0,1405} & \mathbf{0,5960} & \mathbf{0,1328} \\ \mathbf{-0,3454} & \mathbf{0,1328} & \mathbf{0,4869} \end{bmatrix}$$

O vetor b contendo os parâmetros da curva fixa foi obtido pela expressão $b = (f' f)^{-1}(f'\mathbf{y})$, onde **y'= [15.14 26.82 35.72 41.84 45.1]** representa a média de produção aos 12, 30, 48, 66 e 84 meses, respectivamente. A matriz f representa o produto entre a matriz de idades padronizadas (**M**) com a matriz que descreve os três primeiros polinômios de Legendre (L), $f = \mathbf{M}L$, tal que:

$$\mathbf{M} = \begin{bmatrix} \mathbf{1} & \mathbf{-1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-0,5} & \mathbf{0,25} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0,5} & \mathbf{0,25} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix}; \quad L = \begin{bmatrix} \mathbf{0,7071} & \mathbf{0} & \mathbf{0,7906} \\ \mathbf{0} & \mathbf{1,2247} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{2,3717} \end{bmatrix} \text{ e } f = \begin{bmatrix} \mathbf{0,7071} & \mathbf{-1,2247} & \mathbf{1,5811} \\ \mathbf{0,7071} & \mathbf{-0,6123} & \mathbf{-0,1976} \\ \mathbf{0,7071} & \mathbf{0} & \mathbf{-0,7906} \\ \mathbf{0,7071} & \mathbf{0,6123} & \mathbf{-0,1976} \\ \mathbf{0,7071} & \mathbf{1,2247} & \mathbf{1,5811} \end{bmatrix}$$

resultando em $b' = \left((f' f)^{-1}(f'\mathbf{y})\right)' =$ **[47.8832 12.2381 –2.3636]**, que é o vetor de solução para a curva fixa (Figura 1) e comum a todos os indivíduos envolvidos na análise representando o intercepto e os coeficientes linear e quadrático, da equação, respectivamente.

O efeito fixo de local foi criado de forma que cada combinação de progenitores gerasse 10 proles, distribuídas em três locais diferentes, não sendo esperadas diferenças significativas entre os locais.

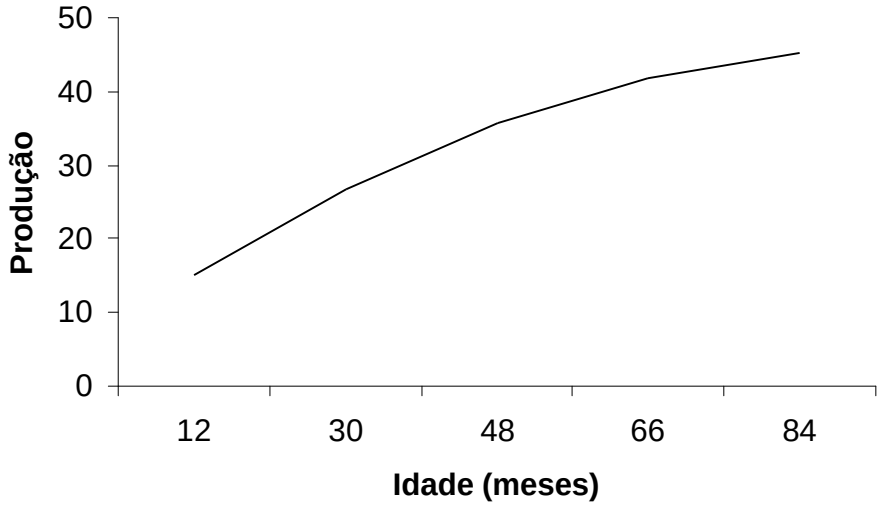

Figura 1- Representação gráfica da regressão da produção em função das idades dos indivíduos, utilizando os três primeiros polinômios ortogonais de Legendre.

Métodos

Métodos de análise

O arquivo de dados foi gerado permitindo que algumas informações fossem perdidas. Portanto, foram escolhidos aleatoriamente 450 indivíduos com informações, os quais tiveram a sua informação referente à produção aos 84 meses de idade deletada, originando um segundo arquivo com 10% de perda de informação. Posteriormente, para os mesmos indivíduos, foram deletadas as informações referentes à produção aos 66, 48 e 30 meses de idade, originando novos arquivos de dados com perda de informação de 20, 30 e 40% de informação, respectivamente. Portanto, os arquivos com informação perdida, representaram a morte de indivíduos em diferentes estágios da trajetória no tempo. A

eliminação da informação desses indivíduos realizada por um processo aleatório, teve por objetivo verificar a eficácia dos métodos comumente utilizados na análise de dados longitudinais, em relação ao efeito de observações perdidas.

Novamente foram escolhidos 450 indivíduos, que tiveram as informações aos 84 meses eliminadas. Posteriormente, para os mesmos indivíduos, foram deletadas as informações referentes à produção aos 66, 48 e 30 meses de idade, originando novos arquivos de dados com perda de informação de 10, 20, 30 e 40% de informação, respectivamente. Porém, neste caso, os 450 indivíduos que tiveram suas informações deletadas, foram escolhidos como dentre aqueles de menor valor fenotípico, representando então amostras sob efeito de seleção. Neste caso, a eliminação da informação nesses indivíduos, realizada com base em um processo seletivo, teve por objetivo verificar a eficácia dos métodos de análises comumente utilizados para análise de dados longitudinais, em relação ao efeito de observações perdidas por meio do efeito de seleção.

Os registros de produção em cada idade, no arquivo de dados completo e nos arquivos com informações perdidas, foram analisados para estimação dos componentes de (co)variância e de parâmetros genéticos, em quatro situações. Na primeira situação, as produções em cada idade foram consideradas características distintas, sendo analisadas por meio de um modelo de característica única. Na segunda situação, as produções em cada idade diferente, foram consideradas medidas repetidas no mesmo indivíduo, onde foi utilizado o modelo de repetibilidade. Na terceira situação, os dados de produção em cada idade foram considerados medidas de dados longitudinais, ou seja, cada idade foi considerada um ponto em uma trajetória ao longo do tempo, sendo empregado neste caso, um modelo de regressão aleatória. Na quarta situação, as produções em cada idade foram consideradas características distintas, sendo analisadas por meio de um modelo de características múltiplas, onde foram processadas análises bicaracter.

Toda a geração e manipulação dos arquivos de dados foi realizada utilizando o aplicativo Statistical Analysis System (SAS, 1990).

Modelos de análise

Modelos de Características Simples (Modelo Unicaracterística)

Para a análise de características simples, as idades foram consideradas como características distintas e foram analisadas separadamente. O modelo utilizado é descrito como:

$$\mathbf{y} = \mathbf{X}b + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

sendo $\mathbf{y}$ o vetor referente à observação da variável resposta; $\mathbf{X}$ é a matriz de incidência de efeitos fixos de local, b é o vetor de soluções para os níveis do efeito fixo de local; $\mathbf{Z}$ é a matriz de incidência do efeito aleatório, $\mathbf{a}$ e $\mathbf{e}$ são os vetores de efeitos aleatórios genético aditivo e residual, respectivamente. As pressuposições a respeito da distribuição dos vetores **y**, **a** e **e**, podem ser descritas como:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{a}\\ \mathbf{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{X}b\\ 0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{ZGZ}' + \mathbf{R} & \mathbf{ZG} & \mathbf{R}\\ \mathbf{GZ}' & \mathbf{G} & 0\\ \mathbf{R} & 0 & \mathbf{R}\end{bmatrix}\right),$$

em que,

$\mathbf{G} = \mathbf{A}\ \sigma^2_a$ e

$\mathbf{R} = \mathbf{I_n}\ \sigma^2_e$

sendo $\mathbf{A}$ a matriz de numeradores dos coeficientes de parentesco entre indivíduos, de ordem igual ao número de indivíduos; σ^2_a é a variância genética aditiva da característica; $\mathbf{I_n}$ é uma matriz identidade, de ordem igual ao número total de observações (**n**); e σ^2_e é a variância residual da característica.

Modelo de Repetibilidade

Para a análise considerando cada idade como medidas repetidas do mesmo indivíduo, o modelo de repetibilidade é descrito como:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wp} + \mathbf{e},$$

em que **y** é um vetor **n** x 1, de **n** observações referentes à produção dos indivíduos, **X** é a matriz de incidência dos efeitos fixos de local e da covariável idade do indivíduo à produção, **Z** e **W** são as matrizes de incidência dos efeitos aleatórios genético aditivo e de ambiente permanente do indivíduo, associados aos vetores **a** e **p**, dos valores genéticos aditivos e de ambiente permanente, respectivamente, e **e** é o vetor de resíduos de mesma dimensão de **y**.

As pressuposições à respeito da distribuição dos vetores **y**, **b**, **a**, **p** e **e** podem ser descritas como:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{Xb} \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{ZGZ}' + \mathbf{WPW}' + \mathbf{R} & \mathbf{ZG} & \mathbf{WP} & \mathbf{R} \\ \mathbf{GZ}' & \mathbf{G} & 0 & 0 \\ \mathbf{PW}' & 0 & \mathbf{P} & 0 \\ \mathbf{R} & 0 & 0 & \mathbf{R} \end{bmatrix} \right),$$

em que $\mathbf{G} = \mathbf{A}\,\sigma^2_a$, $\mathbf{P} = \mathbf{I}_n\,\sigma^2_p$ e $\mathbf{R} = \mathbf{I}_n\,\sigma^2_e$, sendo **A** uma matriz cujos elementos são os numeradores do coeficiente de parentesco entre os indivíduos, de ordem igual ao número de indivíduos (**N**); σ^2_a, σ^2_p e σ^2_e são os componentes de variância associados aos efeitos genético aditivo, de ambiente permanente e de ambiente temporário, $\mathbf{I_n}$ é uma matriz identidade, de ordem igual ao número total de observações (**n**).

Modelo de Regressão Aleatória

O modelo de regressão aleatória utilizado considerou cada idade como um ponto em uma trajetória numa escala contínua, ajustando funções de covariância tanto para o

efeito genético aditivo como para o efeito de ambiente permanente, onde ambas as funções utilizaram os três primeiros polinômios de Legendre, caracterizando uma função polinomial de segundo grau. Este modelo pode ser descrito por:

$$\mathbf{y}_{ij} = \mathbf{F}_{ij} + \sum_{m=0}^{2} b_m f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} a_{im} f_m(\mathbf{a}^*_{ij}) + \sum_{m=0}^{2} g_{im} f_m(\mathbf{a}^*_{ij}) + e_{ij}$$

em que $\mathbf{y_{ij}}$ é a **j**-ésima produção do **i**-ésimo indivíduo; $\mathbf{a}^*_{ij}$ é a idade na produção padronizada entre –1 a +1; f_m é o **m**-ésimo polinômio de Legendre; $\mathbf{F_{ij}}$ é o efeito fixo de local; b_m são os coeficientes de regressão para modelar a trajetória média de todos os indivíduos; a_{im} e g_{im} são os coeficientes de regressão aleatória dos efeitos genético aditivo e de ambiente permanente do indivíduo, respectivamente, e e_{ij} é o efeito do ambiente temporário.

Em notação matricial o modelo de regressão aleatória pode ser representado como:

$$\mathbf{y} = \mathbf{X}b + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{p} + \mathbf{e}$$

sendo **y** o vetor referente a **n** observações de produção em cada idade; **X** é a matriz de incidência de efeito fixo de local e das idades padronizadas entre –1 a +1, que descrevem a trajetória média de todos os indivíduos por meio dos polinômios de Legendre; **ß** é o vetor de soluções do efeito fixo de local e das soluções da regressão média (fixa) de todos indivíduos; **Z** e **W** são matrizes de covariáveis referentes às idades padronizas em cada produção, associadas aos coeficientes de regressão aleatória dos efeitos aleatórios genético aditivo e de ambiente permanente para cada indivíduo, respectivamente, **a** e **p** são vetores contendo os coeficientes de regressão aleatória para cada indivíduo, para os efeitos genético aditivo e de ambiente permanente, respectivamente. O vetor **e** representa os efeitos aleatórios de ambiente temporário. As pressuposições da distribuição dos vetores **a**, **p** e **e**, são:

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} \sim \mathbf{N(0,V)} \text{ com } \mathbf{V} = \begin{bmatrix}\mathbf{A} \otimes \mathbf{Ka} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I} \otimes \mathbf{Kp} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{bmatrix},$$

sendo **A** a matriz de numeradores dos coeficientes de parentesco entre indivíduos, de ordem igual ao número de indivíduos (**N**); **Ka** é a matriz de covariância entre os coeficientes de regressão aleatória do efeito genético aditivo; **Kp** é a matriz de covariância entre os coeficientes de regressão aleatória que descrevem o efeito de ambiente permanente; **I** é uma matriz identidade, de ordem igual ao número total de observações (**n**) e **R** é uma matriz diagonal de variância residual associada a cada observação.

Modelo multi-característica

Para a análise de características múltiplas, as características foram analisadas de duas em duas (análise bicaracterística). O modelo utilizado analisando-se cada idade como uma característica distinta, duas a duas, é descrito como:

$$\mathbf{y} = \mathbf{X}b + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

em que:

$$\mathbf{y} = \begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix}; \mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{f}\\ \mathbf{f} & \mathbf{X}_2\end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}; \mathbf{Z} = \begin{bmatrix}\mathbf{Z}_1 & \mathbf{f}\\ \mathbf{f} & \mathbf{Z}_2\end{bmatrix}; \mathbf{a} = \begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix}; \mathbf{e} = \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}$$

sendo **y** o vetor referente à observação da variável resposta; **X** é a matriz de incidência de efeitos fixos de local, **ß** é o vetor de soluções para os níveis do efeito fixo de local; **Z** é a matriz de incidência do efeito aleatório, **a**, e **e** são os vetores de efeitos aleatórios genético aditivo e residual, respectivamente. As pressuposições a respeito da distribuição dos vetores **y**, **a** e **e**, podem ser descritas como:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{a} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{Xb} \\ \mathbf{f} \\ \mathbf{f} \end{bmatrix}, \begin{bmatrix} \mathbf{ZGZ}' + \mathbf{R} & \mathbf{ZG} & \mathbf{R} \\ \mathbf{GZ}' & \mathbf{G} & \mathbf{f} \\ \mathbf{R} & \mathbf{f} & \mathbf{R} \end{bmatrix} \right),$$

em que

$$\mathbf{G} = \mathbf{A} \otimes \mathbf{G}_0 \text{ e}$$
$$\mathbf{R} = \mathbf{I}_n \otimes \mathbf{R}_0$$

sendo **A** a matriz de numeradores dos coeficientes de parentesco entre indivíduos, de ordem igual ao número de indivíduos; $\mathbf{G_0}$ é a matriz de (co)variância genética aditiva entre as características; $\mathbf{I_n}$ é uma matriz identidade, de ordem igual ao número total de observações (**n**); e $\mathbf{R_0}$ é matriz de variância residual entre as características.

Para a comparação dos resultados obtidos nos diferentes arquivos de dados com as diferentes estratégias de análise, foram avaliadas as estimativas de variâncias e parâmetros genéticos, e em algumas situações foi utilizado o teste da razão de verossimilhança (RAO, 1973). O teste permite comparar dois modelos hierárquicos. Assim, para a comparação do modelo **i**, que contém **n** parâmetros aleatórios, com o modelo **j**, que é o modelo reduzido com **p** parâmetros, a estatística do teste é dada por $\mathbf{LR_{ij}} = \mathbf{2Log_e\, L_i} - \mathbf{2Log_e\, L_j}$, em que $\mathbf{LR_{ij}}$ representa a estatística do teste de razão de verossimilhança para modelos seqüencialmente reduzidos; $\mathbf{Log_eL_i}$ e $\mathbf{Log_eL_j}$ são o logarítmo natural da função de verossimilhança restrita dos modelos **i** e **j**, respectivamente. A estatística do teste apresenta assintoticamente distribuição qui-quadrado com (**n - p**) graus de liberdade. A hipótese de nulidade testada consiste na igualdade dos valores das funções de verossimilhança restrita entre os modelos comparados, H_0:$\mathbf{2Log_eL_i = 2Log_eL_j}$.

Todas as análises foram realizadas pelo aplicativo DFREML Versão 3.0 α (MEYER, 1998).

RESULTADOS E DISCUSSÃO

A distribuição do número de informações (**N**) por idade, estimativas de médias, desvios-padrão (DP), valores mínimo (Min.) e máximo (Max.) e a amplitude da produção em cada idade, considerando-se cada idade como uma característica distinta (modelo unicaracterística), é exibida na Tabela 1. Verificou-se que quando a perda de informação dos dados é realizada por um processo aleatório, as médias estimadas, e a variação dos dados em torno da média se mantêm próximas da situação em que os dados são completos.

Tabela 1: Distribuição do número de observações, média, desvio padrão, mínimo, máximo e amplitude, referentes à produção em cada idade de avaliação, para o conjunto completo de dados, dados incompletos sem seleção e dados incompletos com seleção

Característica	N	Média	DP	Mín.	Máx.	Amplitude
Dados completos						
C1	900	15,10	1,98	9,28	21,51	12,23
C2	900	26,86	1,95	20,16	33,45	13,29
C3	900	35,67	2,19	28,77	42,02	13,25
C4	900	41,87	2,21	35,04	47,72	12,68
C5	900	45,23	2,83	35,62	54,56	18,94
Dados incompleto sem seleção						
C1	900	15,10	1,98	9,28	21,51	12,23
C2	450	26,84	1,95	20,16	33,04	12,88
C3	450	35,66	2,23	29,72	41,57	11,85
C4	450	41,94	2,21	35,04	47,22	12,18
C5	450	45,30	2,74	37,36	54,17	16,81
Dados incompletos com seleção						
C1	900	15,10	1,98	9,28	21,51	12,23
C2	450	27,49	1,92	21,95	33,45	11,50
C3	450	36,92	1,83	32,01	42,02	10,01
C4	450	43,41	1,52	39,57	47,72	8,15
C5	450	47,01	2,22	40,45	54,56	14,11

Por outro lado, a eliminação dos indivíduos de menores valores fenotípicos nas idades mais avançadas, provocou um aumento na estimativa da média, causando, porém, a redução da variabilidade em cada idade e maior semelhança entre os indivíduos (menor amplitude).

As estimativas dos componentes de variância genética aditiva e ambiental, e as estimativas de herdabilidade, obtidas em cada característica para os dados completos, dados com perda de informação ao acaso e dados com perda de informação utilizando a seleção, podem ser observadas na Tabela 2. Verificou-se que quando a perda de informação se deu por um processo aleatório, as estimativas de herdabilidade, de uma forma geral, foram próximas daquelas com os dados completos. No entanto, quando a perda de informação se deu por um processo de seleção, houve uma redução da variância genética aditiva, e conseqüentemente, uma redução nas estimativas de herdabilidade.

Tabela 2 – Estimativas da variância genética aditiva, variância ambiental e herdabilidade para o conjunto de dados completos e para os dados incompletos com e sem seleção analisando-se cada idade como uma característica, separadamente (Modelo unicaracterístico)

	Completo			Sem seleção			Com seleção		
Característica	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_e$	$\hat{h}^2$	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_e$	$\hat{h}^2$	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_e$	$\hat{h}^2$
C1	0,495	3,403	0,13	0,495	3,403	0,13	0,495	3,403	0,13
C2	0,746	3,067	0,20	1,091	2,763	0,28	0,433	3,221	0,12
C3	1,019	3,821	0,21	1,157	3,841	0,23	0,201	3,157	0,06
C4	2,048	2,922	0,41	1,921	3,004	0,39	0,057	2,263	0,02
C5	3,019	4,921	0,38	2,932	4,605	0,39	0,785	4,110	0,16

Quando a análise foi realizada por meio do modelo de repetibilidade, na situação onde houve perda de informação de forma aleatória, as estimativas de herdabilidade foram próximas entre as idades (Tabela 3). Porém quando a perda de informação se deu por meio de seleção, a exemplo do mesmo comportamento das análises unicaracterísticas realizadas em cada idade separadamente, houve uma queda das variâncias, com queda mais acentuada da variância genética aditiva, conforme houve maior perda de informação.

Tabela 3 – Estimativas da variância genética aditiva, variância do efeito de ambiente permanente, variância do efeito de ambiente temporário, variância ambiental e herdabilidade, para a análise utilizando-se o modelo de repetibilidade sem e com seleção

Estimativas dos	Perda de informação sem seleção(%)					Perda de informação com seleção(%)			
Parâmetros	0	10	20	30	40	10	20	30	40
$\hat{s}^2_a$	1,225	1,101	0,994	0,969	0,979	0,970	0,694	0,464	0,299
$\hat{s}^2_p$	0,217	0,271	0,285	0,280	0,311	0,215	0,121	0,058	0,012
$\hat{s}^2_e$	3,666	3,358	3,401	3,398	3,481	3,368	3,357	3,325	3,378
$\hat{s}^2_p + \hat{s}^2_e$	3,883	3,629	3,686	3,678	3,792	3,583	3,478	3,383	3,391
$\hat{h}^2$	0,24	0,23	0,21	0,21	0,21	0,21	0,17	0,12	0,08

As estimativas de variâncias genéticas aditivas em cada idade avaliada, para o conjunto de dados completos, dados com perda de informação sem seleção e com seleção, por meio de modelos de regressão aleatória, estão dispostas na Tabela 4.

Tabela 4 – Estimativas da variância genética aditiva em cada idade para o conjunto de dados completo e para o conjunto de dados incompleto, com e sem seleção, adotando-se o modelo de regressão aleatória

Idade	Dados	Perda de Informação sem seleção(%)				Perda de Informação com seleção(%)			
(Meses)	completos	10	20	30	40	10	20	30	40
12	0,613	0,606	0,613	0,569	0,530	0,596	0,599	0,667	0,580
30	0,683	0,719	0,719	0,827	0,901	0,734	0,789	0,695	0,313
48	1,010	1,080	1,044	1,196	1,320	1,028	0,785	0,480	0,149
66	1,793	1,920	1,818	1,845	1,933	1,683	0,693	0,309	0,176
84	3,304	3,549	3,359	3,031	3,030	2,987	0,784	1,015	0,932

Tabela 5 – Estimativas de variância do ambiente permanente em cada idade e do ambiente temporário ($\hat{s}^2_e$), para o conjunto de dados completo e para o conjunto de dados incompleto, com e sem seleção, adotando-se o modelo de regressão aleatória

Idade	Dados	Perda de Informação sem seleção(%)				Perda de Informação com seleção(%)			
(Meses)	completos	10	20	30	40	10	20	30	40
12	1,252	1,251	1,206	1,300	1,372	1,231	1,163	1,294	1,383
30	1,079	1,080	1,056	1,042	1,038	1,113	1,168	0,978	1,009
48	1,598	1,577	1,551	1,549	1,453	1,651	1,515	1,235	1,214
66	1,228	1,070	1,100	1,144	1,061	1,051	0,507	0,493	0,545
84	2,630	2,039	2,164	2,464	2,438	1,834	1,863	1,943	1,906
$\hat{s}^2_e$	2,062	2,054	2,097	2,017	2,004	2,085	2,142	1,999	1,983

Tabela 6 – Estimativas de herdabilidade em cada idade para o conjunto de dados completo e para o conjunto de dados incompleto, com e sem seleção, adotando-se o modelo de regressão aleatória

Idade	Dados	Perda de Informação sem seleção(%)				Perda de Informação com seleção(%)			
(Meses)	completos	10	20	30	40	10	20	30	40
12	0,16	0,15	0,15	0,15	0,14	0,15	0,15	0,17	0,15
30	0,18	0,18	0,18	0,21	0,23	0,19	0,19	0,19	0,09
48	0,22	0,23	0,22	0,25	0,28	0,22	0,18	0,13	0,04
66	0,35	0,38	0,36	0,37	0,39	0,35	0,21	0,11	0,06
84	0,41	0,46	0,44	0,40	0,41	0,43	0,16	0,20	0,19

Verificou-se que a perda de informação por um processo ao acaso, praticamente não alterou as estimativas de herdabilidade em qualquer nível de desbalanceamento (Tabela 6). De acordo com DAL ZOTTO (2000), em modelos de regressão aleatória o número mínimo de observações a ser considerado para cada nível de efeito de ambiente permanente, deve ser igual a um mais o número de parâmetros da função utilizada para descrever a trajetória dos dados. Por outro lado, segundo SCHAEFFER e DEKKERS (1994), modelos de regressão aleatória permitem o uso de indivíduos com somente uma observação, e os resultados observados neste estudo foram condizentes com esses autores.

Quando a perda de informação se deu pelo descarte dos indivíduos de menores valores fenotípicos, as estimativas de variâncias genéticas e ambientais diminuíram, ocasionando em menores estimativas de herdabilidade. Sob seleção, somente as estimativas de herdabilidade do modelo de regressão aleatória com 10% de perda de informação foram semelhantes às estimativas obtidas utilizando o conjunto de dados completo. RESENDE et al. (2001), aplicando modelos de regressão aleatória para descrever o diâmetro à altura do peito de um aos sete anos de idade em *Eucalyptus urophylla*, verificaram que as estimativas de herdabilidade foram próximas das obtidas por meio de modelos unicaracterística até os três anos de idade. Para as idades mais avançadas, as estimativas foram menores. Os autores discutiram que análises por meio de modelos unicaracterísticos, superestimaram parâmetros por causa da redução da variância fenotípica, devido à morte de indivíduos menos vigorosos, tratando-se então de uma população naturalmente selecionada para adaptação. MATHESON E RAYMOND (1984), citados por RESENDE et al. (2001), encontraram estimativas de herdabilidade em duas populações de *Pinus radiata* iguais a 0,12 e 0,24. As estimativas assumiram os valores de 0,21 e 0,33 após a eliminação das piores plantas, respectivamente. Estes resultados foram contrários aos obtidos neste estudo, provavelmente pelo tamanho da amostra, pois os dados foram gerados de forma a apresentar um teste de progênie, contendo somente duas gerações, a geração parental e a geração das proles. Assim, a eliminação de informação das plantas com menores valores fenotípicos e conseqüentemente, a eliminação de plantas de menores valores genotípicos afetou de forma acentuada a estrutura dos dados.

Os dados completos foram analisados por meio de dois modelos de regressão aleatória, onde a diferença entre eles consistiu na pressuposição da variância do efeito de ambiente temporário. No primeiro modelo as variâncias foram consideradas constantes, e no segundo, cada idade apresentou uma variância diferente. Comparando as estimativas de componentes de variâncias, de herdabilidade, e ainda, comparando os modelos por meio dos valores da função de verossimilhança (Tabela 7), verificou-se a aceitação da mesma, para um nível de 5% de significância, indicando que neste caso, considerar as

variâncias do efeito de ambiente temporário como homogêneas ou heterogêneas, não alterou o ajuste da função.

Tabela 7 – Estimativas de variâncias genética aditiva (σ_a^2), de ambiente permanente (σ_p^2), e de ambiente temporário (σ_e^2), e de herdabilidade (h^2), em cada idade, obtidas por meio de modelos de regressão aleatória sob diferentes pressuposições, valores da função de verossimilhança e do teste da razão de verossimilhança(c^2)

	Resíduo homogêneo				Resíduo heterogêneo			
Idades	σ_a^2	σ_p^2	σ_e^2	h^2	σ_a^2	σ_p^2	σ_e^2	h^2
12	0,623	1,236	2,067	0,16	0,586	1,263	2,043	0,15
30	0,705	1,048	2,067	0,18	0,707	1,031	2,084	0,18
48	1,089	1,537	2,067	0,23	1,116	1,532	2,113	0,23
60	1,811	1,223	2,067	0,35	1,876	1,198	1,918	0,38
84	2,916	2,864	2,067	0,37	3,083	2,523	2,389	0,39
-2Log(L)		10624,23				10621,64		
c^2	2,59 (P > 0,05)							

Quando os dados foram analisados considerando cada idade como uma característica, por meio de modelos multi-característica, sendo analisadas as características aos pares (modelo bicaracter), verificou-se o mesmo comportamento dos resultados anteriores. Na Tabela 8 estão representadas as estimativas de herdabilidade para os dados completos, e dados incompletos, com e sem seleção. Esses valores representam as estimativas mínima e máxima, obtidas analisando-se as características (idades) duas a duas. A seleção provocou redução da variabilidade, à medida que se aumentou a perda de informação, ou seja, este modelo mostrou-se bastante sensível ao efeito da seleção. Entretanto, todas as estimativas de herdabilidade foram próximas quando a perda de informação não se deu por processo seletivo.

De uma forma geral, independente de como os dados foram analisados, seja por modelos unicaracterística (unicaracter e de repetibilidade), de multi-característica ou por regressão aleatória (admitindo que as idades representam pontos em uma escala contínua), o efeito da seleção provocou menor discriminação entre os indivíduos, diminuindo a

variabilidade entre os mesmos. Como resultado dessa menor variabilidade entre indivíduos, a variância genética aditiva diminuiu, resultando em menores estimativas de herdabilidade. Com 10% de perda de informação, praticamente não houve alteração nos valores de herdabilidade; com 20% houve redução nos valores de herdabilidade nas duas últimas idades; com 30% a redução nos valores de herdabilidade ocorreu nas três últimas idades e com 40% houve redução nas quatro últimas idades. Porém, com 10% de perda de informação sob seleção, o modelo de regressão aleatória foi superior aos demais.

Tabela 8: Estimativas de herdabilidade obtidas para o conjunto de dados completo e para o conjunto de dados incompleto com e sem seleção, adotando-se o modelo multi-característica em análises bicaracter

		Perda de informação sem seleção(%)			
Idades	Completo	10%	20%	30%	40%
12	0,12 - 0,14	0,13 - 0,14	0,13 - 0,14	0,11 - 0,13	0,12 - 0,13
30	0,20 - 0,21	0,20 - 0,25	0,20 - 0,25	0,20 - 0,25	0,27 - 0,29
48	0,19 - 0,23	0,19 - 0,23	0,19 - 0,23	0,20 - 0,26	0,23 - 0,26
66	0,39 - 0,42	0,33 - 0,40	0,38 - 0,42	0,38 - 0,42	0,38 - 0,40
84	0,37 - 0,40	0,32 - 0,43	0,38 - 0,41	0,38 - 0,41	0,38 - 0,40

		Perda de informação com seleção(%)			
Idades	Completo	10%	20%	30%	40%
12	0,12 - 0,14	0,13 - 0,14	0,13 - 0,15	0,13	0,13
30	0,20 - 0,21	0,20 - 0,25	0,19 - 0,20	0,20 - 0,25	0,12 - 0,14
48	0,19 - 0,23	0,19 - 0,23	0,19 - 0,22	0,05 - 0,07	0,05 - 0,06
66	0,39 - 0,42	0,40 - 0,42	0,03 - 0,11	0,03 - 0,07	0,03 - 0,06
84	0,37 - 0,40	0,11 - 0,23	0,11 - 0,18	0,15 - 0,19	0,16 - 0,18

Suspeitando que a perda de informação poderia alterar a variância do efeito de ambiente temporário, os dados foram analisados novamente com 40% de perda e sob seleção, porém admitindo variâncias heterogêneas para o efeito de ambiente temporário em cada idade. Além disso, os dados completos foram também analisados para verificação da homogeneidade de variâncias do efeito de ambiente permanente. Comparando as estimativas de componentes de variâncias, de herdabilidade, e ainda,

comparando os modelos por meio dos valores da função de verossimilhança (Tabela 9), verificou-se que considerar a heterogeneidade de variância, tanto para os dados completos, como para os dados selecionados, não promoveu melhorias de ajuste na descrição da variação dos dados. Portanto, a perda de informações não alterou a estrutura da variância do efeito de ambiente temporário.

Tabela 9 – Estimativas de variâncias genética aditiva ($\hat{s}^2_a$), de ambiente permanente ($\hat{s}^2_p$), de ambiente temporário ($\hat{s}^2_e$), de herdabilidade ($\hat{\mathbf{h}}^2$) em cada idade, obtidas por meio de modelos de regressão aleatória sob diferentes pressuposições; valores do logarítmo da função de verossimilhança e do teste da razão de verossimilhança(c^2) para os dados completos e selecionados com 40% de perda de informação

Dados Completos

	Resíduo homogêneo				Resíduo heterogêneo			
Idades	$\hat{s}^2_a$	$\hat{s}^2_p$	$\hat{s}^2_e$	$\hat{\mathbf{h}}^2$	$\hat{s}^2_a$	$\hat{s}^2_p$	$\hat{s}^2_e$	$\hat{\mathbf{h}}^2$
12	0,613	1,252	2,062	0,16	0,618	1,261	2,044	0,16
30	0,683	1,079	2,062	0,18	0,702	1,049	2,090	0,18
48	1,010	1,598	2,062	0,22	1,044	1,561	2,154	0,22
60	1,793	1,228	2,062	0,35	1,825	1,187	1,890	0,37
84	3,304	2,630	2,062	0,41	3,277	2,337	2,480	0,40
-2Log(L)		10624,22				10621,64		
c^2	2,58 (P > 0,05)							

Dados Selecionados

Idades	$\hat{s}^2_a$	$\hat{s}^2_p$	$\hat{s}^2_e$	$\hat{\mathbf{h}}^2$	$\hat{s}^2_a$	$\hat{s}^2_p$	$\hat{s}^2_e$	$\hat{\mathbf{h}}^2$
12	0,580	1,383	1,984	0,15	0,565	1,509	1,838	0,14
30	0,313	1,009	1,984	0,09	0,298	0,993	2,230	0,08
48	0,149	1,214	1,984	0,04	0,141	1,215	1,963	0,04
60	0,176	0,545	1,984	0,06	0,184	0,517	1,693	0,07
84	0,932	1,906	1,984	0,19	0,920	1,495	2,588	0,18
-2Log(L)		5979,28				5970,94		
c^2	8,34 (P > 0,05)							

As estimativas de herdabilidade provenientes dos modelos de regressão aleatória, empregados na situação em que os dados foram completos, ou ainda incompletos sem e

com o processo de seleção, se aproximaram das estimativas provenientes da análise de modelos multi-característica. Porém, uma análise nas estimativas de herdabilidade mostraram que no modelo multi-característica essas estimativas foram mais afetadas pelo desbalanceamento do que aquelas obtidas por meio do modelo de regressão aleatória. Com 10% de perda de informação, por exemplo, as estimativas de herdabilidade obtidas por meio do modelo de regressão aleatória foram bem próximas daquelas obtidas com os dados completos. Por outro lado, o modelo multi-característica, no mesmo nível de desbalanceamento, apresentou estimativas bastante subestimadas na última idade mensurada. Portanto, como os modelos de regressão aleatória utilizam funções de covariância que fornecem uma descrição contínua da estrutura de covariância dos efeitos aletórios associados ao caráter analisado, pode-se afirmar que os mesmos podem expressar de maneira mais correta, a variância dos efeitos aleatórios do modelo linear misto que procuram descrever dados de natureza longitudinal. Por considerar que o caracter em questão pode sofrer variações ao longo do tempo, torna-se uma técnica mais realística do que modelo de repetibilidade, e por utilizar menor número de parâmetros, torna-se mais atrativo que a utilização de modelos multi-característica, que por sua vez, podem ser proibitivos na prática quando envolvem um grande número de parâmetros.

SCHAEFFER e WILTON (1998) discutiram que em situações onde todas as características são observadas em cada indivíduo, quando as herdabilidades das características são próximas, e todas as características são positivamente correlacionadas, a análise com o modelo que considera múltiplas características, poderá não fornecer um incremento significativo no valor da acurácia das avaliações genéticas.

Ao analisar o padrão de correlação genética entre as idades avaliadas, verificou-se que o modelo de regressão aleatória que utilizou os dados completos se aproximou bem dos valores obtidos pelo modelo com 10% de perda de informação (Figura 1a e 1b). Por outro lado, quando os dados foram mais influenciados pela seleção (acima de 20%), ocorreu uma diminuição da correlação entre as idades, principalmente entre as idades mais distantes (Figura 1c, 1d, e 1e).

HENDERSON (1975), no contexto do melhoramento genético de medidas que se repetem no tempo, discutiu que as predições do mérito genético dos indivíduos seriam não-viesadas pela seleção se as variâncias e covariâncias fossem conhecidas, se a seleção ocorresse dentro de níveis de efeitos fixos e se todos os registros da primeira produção de todos os indivíduos estivessem disponíveis para serem utilizados. LOFGREN et al. (1983), ao analisarem rebanhos de bovinos leiteiros, relataram que o aumento da taxa de descarte de animais pode levar a uma alteração na classificação dos melhores animais. Por outro lado, CASSELL et al. (1983) avaliando o efeito do impacto do descarte de animais sobre a avaliação genética de touros, considerando os registros da primeira e segunda lactação como sendo características distintas (multi-característica), verificaram que as avaliações dos animais sob o modelo multi-característica permaneceram relativamente inalteradas com o aumento no percentual de descarte de animais baseados na segunda lactação. POLLAK et al. (1984), afirmaram que em alguns casos, a adoção de modelos multi-característica pode reduzir ou eliminar o viés devido a seleção.

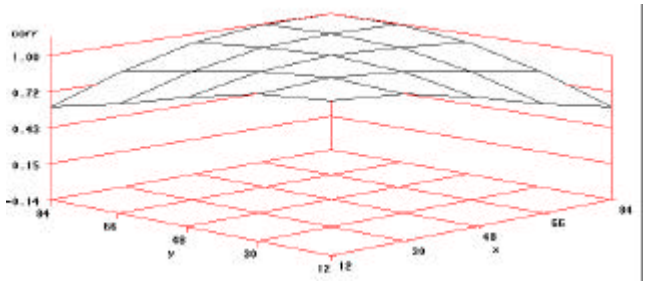

Figura 1 a

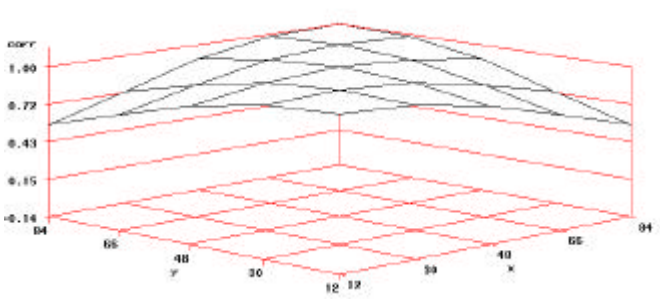

Figura 1 b

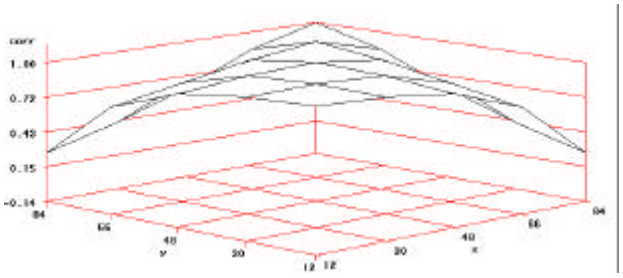

Figura 1 c

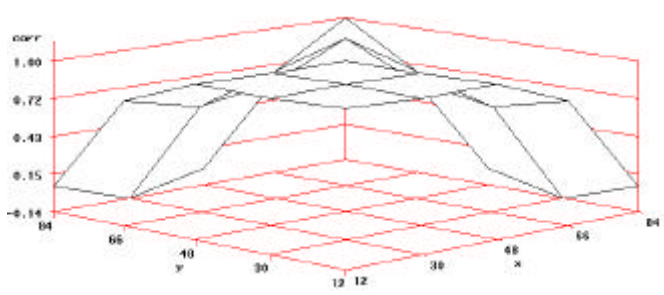

Figura 1 d

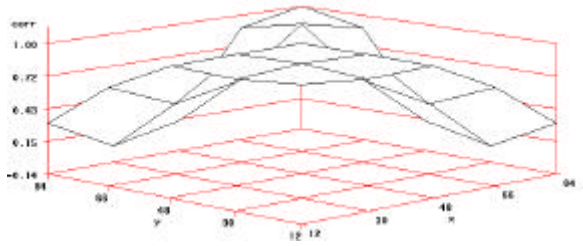

Figura 1 e

Figura 1 – Correlações genéticas entre as idades e suas estimativas obtidas pelos modelos de regressão aleatória completo (a), com perda de informação de 10%(b), com perda de informação de 20%(c), com perda de informação de 30%(d) e com perda de informação de 40%(e).

CONCLUSÕES

Para análise de dados de natureza longitudinal, quando a pressuposição de que a correlação entre mensurações sucessivas no mesmo indivíduo é igual a um não é verdadeira, modelos de regressão aleatória devem ser preferidos por permitirem descrever continuamente as estruturas de covariâncias de crescimento ao longo do tempo, e ainda, por permitirem a estimação de parâmetros genéticos nos vários pontos da trajetória que descreve o crescimento.

Em ausência de seleção, modelos de regressão aleatória permitem a obtenção de estimativas de covariâncias e de parâmetros genéticos muito próximas daquelas obtidas por meio de modelos de multi-características, porém com menor parametrização, o que na prática é uma grande vantagem.

Em amostras populacionais submetidas ao efeito da seleção, caracterizando perdas de informação (desbalanceamento), modelos multi-característica foram mais susceptíveis ao viés da seleção do que análise por meio de modelos de regressão aleatória. Além disso, por considerar a natureza contínua da variável resposta, modelos de regressão aleatória devem ser preferidos aos modelos unicaracterísticos quando os dados apresentam alto grau de desbalanceamento.

REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO, C. V. 2003. Modelos de Regressão Aleatória para avaliação Genética da produção de leite na raça Holandesa. Viçosa. UFV, 2003, 85p. Tese (Doutorado em Zootecnia) – Universidade Federal de Viçosa, M.G., 2003.

CASSEL, B.G., McDANIEL, B.T., ROBISON, O.W. 1983. Impact of culling on sire evaluation by mixed model procedures. *J. Dairy Sci.*, 66:1696-1706.

DAL ZOTTO, R. 2000. Compararison of different test-day models for genetic evaluation of Italian Brown dairy cattle. Proc. Int. Workshop on Genetic Improvement of functional traits in cattle. Interbull Bull.25:95-98.

HENDERSON, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 31:423.

HENDERSON JUNIOR, C. R., 1982. Analysis of covariance in the mixed model: higher level, nonhomogeneus, and random regressions. *Biometrics* v.38, p.623-640, 1982.

HUISMAN, A. E., VEERKAMP, R.F., VAN ARENDONK, J.A.M. 2001 Genetic parameters for different random regression models to describe weight data of pigs. In Annual Meeting of the EAAP, 52. Budapest, Hungria.

KETTUNEN, A., MÄNTYSAARI, E.A., POSO, J. 2000. Estimation of genetic parameters for daily milk yield of primiparous Ayrshire cows by random regression test-day models. *Livest. Prod. Sci.*, 66:251-261.

KIRKPATRICK, M.; LOFSVOLD, D.; BULMER, M. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, vol. 24, n. 3, p. 979-993.

LOFGREN, D.L., CASSELL, B.G., NORMAN, H.D. et al. 1983. Effects culling on sire evaluation by mixed models. *J. Dairy Sci.,* 66(11):2418-2425.

MEYER, K. DXMRR – A set programs to estimate COVARIANCE FUNCTIONS FOR LONGITUDINAL DATA BY REML. In: World Congress of Genetics Applied to Livestock Production, 6, 1998, Armidale. Proceeding… Armidale: University of New England, 1998. CD-ROM.

RAO, C.R. 1973. Linear statistical inference and its applications. 2ed. New York:John Wiley & Sons. 552p.

REKAYA, R., CARABAÑO, M.J., TORO, M.A. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. *Livest. Prod. Sci.*, 57:203-217.

RESENDE, M. D. V.; REZENDE, G. D. S. P.; FERNANDES, J. S. C. 2001. Regressão aleatória e funções de covariância na análises de medidas repetidas. *Revista de Matemática e Estatística*, vol.19:21-40.

SAS INSTITUTE INC. SAS/STAT® user's guide, version 6. 4ed. Carry, NC. 1990. v.1, 943p.

SCHAEFFER, L.R., DEKKERS, J. C. M. Random regression in animal models for test day production in dairy cattle. In: World congress genetic applied livestock production, 5., 1994, Guelph. ON, Canada, *Proccedings*… Guelph, 1994. p. 443-446.

SCHAEFFER, L.R., WILTON, J.W. 1998. Comparison of single and multiple trait beef sire evaluation. *Can. J. Anim. Sci.,* 76:2303-2307.

POLLAK, E.J., VAN DER WERF, J.; QUAAS, R.L. 1984. Selection bias and multiple trait evaluation. *J. Dairy Sci.,* 67: 1590-1595.

RESUMO E CONCLUSÕES

Neste estudo foram utilizados dados simulados de característica longitudinal, referentes à um teste de progênie do cruzamento entre 30 genitores masculinos com 3 genitores femininos diferentes cada um, originando em cada cruzamento 10 indivíduos, distribuídos em três locais diferentes. Foram gerados 4.500 registros utilizados em análises de modelo de regressão aleatória, para verificar o efeito de se assumir diferentes pressuposições à respeito da heterogeneidade de variância ambiental sobre os parâmetros genéticos e componentes de variância. Dois modelos de regressão aleatória, ajustando funções de covariância para o efeito genético aditivo e para o efeito de ambiente permanente, ou ajustando uma função de covariância somente para o efeito genético aditivo, foram utilizados para análises, considerando ou não a heterogeneidade de variâncias do efeito de ambiente temporário, resultando em quatro diferentes modelos de regressão aleatória. Para o ajuste da função de covariância, foram utilizados os polinômios de Legendre de terceira ordem (ajuste de segundo grau). Além disso, foram avaliadas diferentes estratégias de análise de dados longitudinais, frente a diferentes níveis de perda de informação. A eliminação da informação dos indivíduos realizada por um processo aleatório, teve por objetivo verificar a eficácia dos métodos de análise comumente utilizados para a análise de dados longitudinais em relação ao efeito de observações perdidas. Todas as análises foram processadas no aplicativo DFREML, Versão 3.0 α (MEYER, 1998). Com base nos testes da razão de verossimilhança restrita e Critério de informação de Akaike, o melhor modelo foi aquele que considerou como sendo heterogênea, tanto a variância do efeito de ambiente permanente como a variância do efeito de ambiente temporário. Modelos que consideram o efeito de ambiente temporário foram preferidos em relação aos modelos que consideraram a variância do efeito de ambiente permanente. Apesar de os testes indicarem o Modelo 3 (modelo que considerou a variância do efeito de ambiente temporário heterogênea e a do efeito de ambiente permanente constante), o segundo melhor modelo, as

estimativas de herdabilidade do Modelo 2 (que considerou a variância do efeito de ambiente temporário constante e a variância do efeito de ambiente permanente heterogênea) foram mais próximas das estimativas do Modelo 1. Análises obtidas utilizando-se dados com perda de informação sem o efeito da seleção, apresentaram estimativas de herdabilidade próximas daquelas obtidas utilizando-se dados completos. Porém, sob o efeito de seleção, as estimativas de herdabilidade são subestimadas em todos os modelos, principalmente com o aumento do nível de desbalanceamento. Considerar a estrutura de variância ambiental é fundamental para a obtenção de parâmetros genéticos e componentes de variância mais precisos. Quanto maior a variância de um efeito para a característica, maior a importância de se considerar esse efeito quando se utiliza um modelo de regressão aleatória. É evidente, que por se tratar de um estudo de simulação, estudos adicionais são necessários para se avaliar a adoção de diferentes tamanhos de amostra.