

ESTIMAÇÃO DE MODELOS LINEARES GERAIS MISTOS UTILIZANDO O SAS®

O Modelo Linear Geral Misto (MLGM) enquadra-se numa classe de modelos que tem sido tradicionalmente analisada através de procedimentos de análise de variância. Nos MLGM existem três aspectos fundamentais: estimação e significância dos efeitos fixos, predição dos efeitos aleatórios e estimação das componentes de variância. Na análise de MLGM desbalanceados, a estimação das componentes de variância tem importância extrema e depende da estrutura de covariância e dos métodos de estimação utilizados. Este artigo pretende apresentar os principais métodos de estimação do MLGM com estruturas gerais de covariância dos efeitos aleatórios, disponíveis no procedimento "proc mixed" do *Statistical Analysis System (SAS)*.

Luís Pereira – ESGHT

Lara Ferreira – ESGHT¹

INTRODUÇÃO

Os MLGM são utilizados para descrever um conjunto de dados cuja estrutura de tratamentos envolve alguns factores que são fixos e outros que são aleatórios, ou seja, modelos lineares que contêm efeitos fixos e aleatórios, independentemente da média e do erro. Consequentemente, a análise de um modelo misto envolve duas partes: uma análise para a parte aleatória e outra para a parte fixa. Nos MLGM a análise da parte aleatória consiste na predição dos efeitos aleatórios e na estimação das componentes de variância. A análise da parte fixa consiste na estimação e na realização de testes de hipóteses sobre funções estimáveis dos efeitos fixos. Em geral, tanto a predição dos

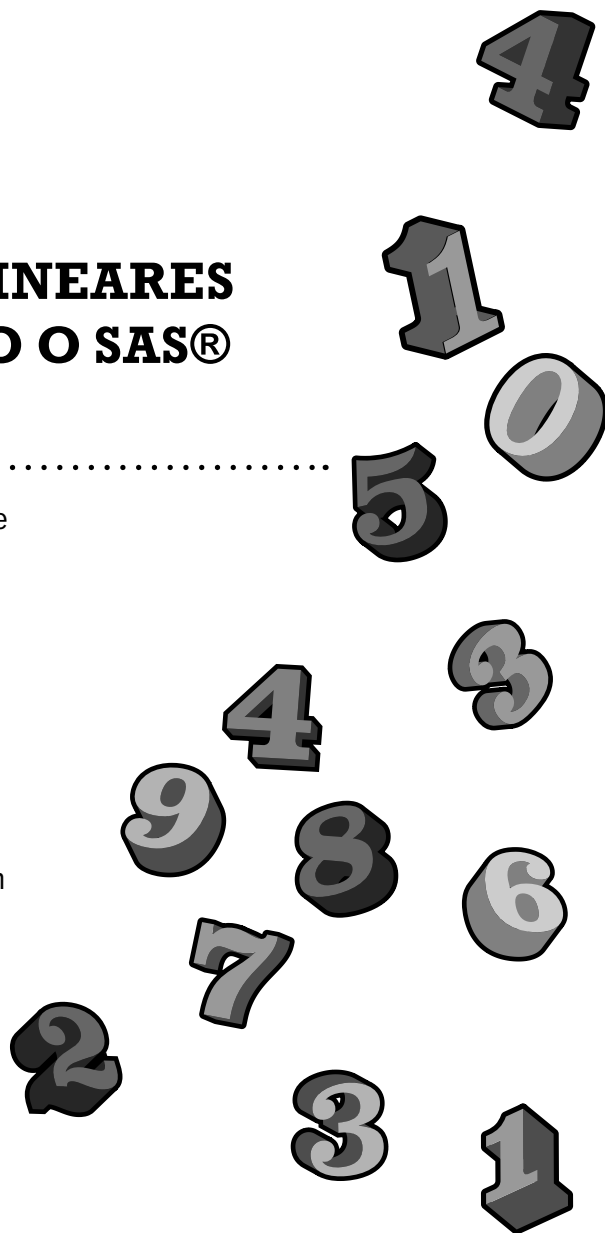
efeitos aleatórios como a estimação dos efeitos fixos dependem da estimação das componentes de variância.

Segundo Scheffé (1959), o MLGM foi amplamente estudado por Fisher em 1918, com grande repercussão nos estudos de genética quantitativa. Tal modelo foi denominado pelo autor como modelo de componentes de variância.

Diversos métodos têm sido propostos para estimar as componentes de variância, destacando-se o método da máxima verosimilhança (*Maximum Likelihood*: ML) da autoria de Hartley e Rao (1967), o método da estimação quadrática não enviesada de variância mínima (*Minimum Variance Quadratic Unbiased Estimation*: MIVQUE) descrito em Rao (1971) e o método da máxima verosimilhança restrita (*Restricted Maximum Likelihood*: REML) descrito por

Patterson e Thompson (1971). Todos estes métodos estão disponíveis no procedimento "proc mixed" do SAS.

Searle et al. (1992) apresentaram uma ampla discussão sobre estimação de componentes de variância e análise de modelos mistos, ilustrando-os por meio de exemplos. Richardson e Welsh (1995) apresentaram duas definições de máxima verosimilhança robusta e máxima verosimilhança restrita robusta, e apresentaram também, por meio de simulação, um estudo para investigar as propriedades assintóticas e as vantagens de utilizar esses métodos robustos. Gilmour et al. (1995) descreveram e aplicaram o algoritmo *Average Information (AI)*, na estimação de componentes de variância pelo método da REML, em modelos mistos com erros correlacionados.



Para além destes métodos, têm sido utilizados no passado recente outros métodos para a estimação das componentes de variância no contexto do MLGM. Destaca-se o método de ajustamento de constantes (Henderson, 1953; Fuller e Battese, 1973). Existe, no entanto, evidência empírica (Swallow e Monaham, 1984) que favorece a utilização dos métodos de verosimilhança. O principal obstáculo à sua utilização reside no grande peso computacional que poderá impedir a convergência ao trabalhar com conjuntos de dados de grande dimensão.

Este trabalho tem como objectivos apresentar métodos de estimação das componentes de variância disponíveis no procedimento "proc mixed" do sistema computacional SAS versão 9.0 e ilustrar a utilização desse procedimento na estimação de MLGM com dois factores desbalanceados.

METODOLOGIA

Considerações iniciais

Neste estudo adopta-se a forma matricial para apresentar o MLGM, descrita em Searle (1987), Searle et al. (1992), Littell et al. (2006), entre outros:

$$y = X\beta + Zu + \epsilon, \quad (1)$$

onde y é o vector das observações da variável dependente de dimensão $(nx1)$; X é a matriz não estocástica de n valores de p variáveis explicativas, de dimensão $(n \times p)$; β é o vector dos efeitos fixos, desconhecido, de dimensão $(p \times 1)$; Z é a matriz de desenho dos efeitos aleatórios, conhecida, de dimensão $(n \times q)$; u é o vector de efeitos aleatórios, desconhecido, de dimensão $(qx1)$ e ϵ é o vector de erros aleatórios não observáveis, de dimensão $(nx1)$.

Assim, para o modelo misto (1), assumindo que os efeitos aleatórios u e ϵ têm distribuição normal, com média zero, e matrizes de variâncias e covariâncias G e $R = \sigma^2 I$ respectivamente, o vector y terá distribuição normal multivariada, com média $X\beta$ e matriz de variâncias e covariâncias $V = ZGZ' + \sigma^2 I$, ou seja, $y \sim N(X\beta, V)$.

Em geral, a matriz de variâncias e covariâncias de y é uma função linear de parâmetros desconhecidos a serem estimados. Assim, quanto mais adequada for a matriz G , escolhida a priori, mais os resultados das estimativas de β e u se aproximarão de soluções BLUE (*Best Linear Unbiased Estimator*) e BLUP (*Best Linear Unbiased Predictor*), respectivamente.

Estimação dos efeitos fixos

Na análise do MLGM tem-se, em geral, interesse na estimação e testes de hipóteses dos efeitos fixos. Entretanto, para a estimativa de uma função estimável dos parâmetros de efeitos fixos é necessário o conhecimento das estimativas das componentes de variância. Assim, as estimativas dos parâmetros de efeitos fixos dependem dos métodos utilizados na obtenção das estimativas das componentes de variância. O Método dos Mínimos Quadrados Generalizados (MMQG), que minimiza

$(y - X\beta)'V^{-1}(y - X\beta)$ fornece o sistema de equações normais generalizadas .

$X'V^{-1}X\beta = X'V^{-1}y$ Assim, para o modelo (1) tem-se que o BLUE de β é dado por:

$$\beta(\hat{\theta}) = (X'V^{-1}X)^{-1} X'V^{-1}y, \quad (2)$$

onde θ é um vector de parâmetros formado pelas componentes de variância e $(X'V^{-1}X)^{-1}$ é uma qualquer matriz inversa generalizada de $X'V^{-1}X$ (Searle, 1971). Desse modo, as componentes de variância são consideradas nas estimativas dos efeitos fixos. Contudo, nas aplicações práticas, as componentes de variância são desconhecidas. Nesses casos, uma estratégia interessante e conveniente consiste em obter estimativas das componentes de variância, as quais são utilizadas em vez das componentes em V . Substituindo θ por $\hat{\theta}$ na expressão (2) tem-se que:

$$\hat{\beta}(\hat{\theta}) = (X'V^{-1}(\hat{\theta})X)^{-1} X'V^{-1}(\hat{\theta})y. \quad (3)$$

Predição dos efeitos aleatórios

No MLGM (1) o vector u é um vector de variáveis aleatórias. Uma questão que se coloca frequentemente é a se-

guinte: dado o vector de dados y , como é que se pode prever os valores dos efeitos aleatórios que poderiam a ele estar associados, ou seja, qual é "um" estimador da média condicional $E(u|y)$? No caso do MLGM (1) tem-se que o melhor preditor linear de u sob a normalidade é dado por:

$$\tilde{u} = E(u|y) = GZ'V^{-1}(y - X\hat{\beta}). \quad (4)$$

Combinando os conceitos de predição e de melhor estimador linear não enviesado, tem-se que o melhor preditor linear não enviesado, ou seja, o BLUP de u , é dado por:

$$\hat{u} = GZ'V^{-1}(y - X\hat{\beta}). \quad (5)$$

Quando os parâmetros de θ são conhecidos, $\hat{\beta}$ é o melhor estimador linear centrado de β e \hat{u} é o melhor preditor linear centrado de u (a este respeito ver Searle, 1971; Harville, 1990; Robinson, 1991, McLean, Sanders e Stroup, 1991). Quando as componentes de variância são conhecidas, o cálculo do BLUP de u não apresenta dificuldades e pode ser obtido através das conhecidas equações normais (Henderson, 1984):

$$\begin{bmatrix} X'X & X'Y \\ Z'XZ' & Z'Y \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix}. \quad (6)$$

Resolvendo-se esse sistema de equações, obtêm-se as soluções para os efeitos fixos $\hat{\beta}$, e as predições para os efeitos aleatórios, \hat{u} . Para o desenvolvimento precedente, assume-se que V é conhecida. Quando não é esse o caso, então essas variâncias devem ser estimadas utilizando-se um dos métodos disponíveis na literatura, como por exemplo o método da ML ou da REML. Então, substituindo V por \hat{V} , tem-se que as soluções das equações normais são dadas por:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} (X'\hat{V}^{-1})X' & X'\hat{V}^{-1}y \\ \hat{G}Z'\hat{V}^{-1}(y - X\hat{\beta}) \end{bmatrix}. \quad (7)$$

Assim, utilizando-se a expressão (7) obtêm-se as estimativas dos efeitos fixos e as predições dos efeitos aleatórios, $\hat{\beta}$ e \hat{u} , respectivamente. Um aspecto interessante das equações normais é que elas podem ser utilizadas em procedimentos iterativos para os cálculos das estimativas ML e REML das componentes de variância. Alguns

$K'y \sim N(\phi, K)$

detalhes do relacionamento entre essas equações e as estimativas ML e REML são apresentados em Harville (1977) e Searle et al. (1992).

Estimação das componentes de variância

Para a obtenção do BLUE de β e do BLUP de u , exige-se o conhecimento das estimativas dos componentes de variância. Um problema relacionado com a estimação das componentes de variância para dados desbalanceados assenta na dificuldade de escolher um método de estimação entre os muitos métodos de estimação disponíveis. Apresentam-se de seguida alguns métodos de estimação das componentes de variância disponíveis no procedimento "proc mixed" do SAS: ML, REML, MIVQUEO.

Método da Máxima Verosimilhança (ML)

Hartley e Rao (1967) aplicaram o método da ML ao MLGM. Este método consiste em maximizar a função de verosimilhança, em relação aos efeitos fixos e às componentes de variância. Assim, para o MLGM (1), assumindo $y \sim N(X\beta, V)$ com $V = ZGZ' + \sigma^2 I$, a função de verosimilhança é dada por:

$$L = (2\pi)^{-\frac{n}{2}} |V|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (y - X\beta)' V^{-1} (y - X\beta)\right], \quad (8)$$

onde $|V|$ é o determinante de V . O logaritmo da função de verosimilhança é dado por:

$$-2 \log L = n \log(2\pi) + \log |V| + (y - X\hat{\beta})' V^{-1} (y - X\hat{\beta}), \quad (9)$$

O "proc mixed" implementa o método ML através da construção de uma função objectivo para ML minimizando $-2l$, ou seja, minimizando:

$$-2 \log L = n \log(2\pi) + \log |V| + (y - X\beta)' V^{-1} (y - X\beta) \quad (10)$$

Minimizando a expressão (10) relativamente a β obtém-se:

$$\hat{\beta} = (X' \tilde{V}^{-1} X)^{-1} X' \tilde{V}^{-1} y, \quad (11)$$

onde $\tilde{\beta}$ e \tilde{V} representam as estimativas ML de β e V , respectivamente. Substituindo $\hat{\beta}$ na expressão (10), tem-se que:

$$-2 \log L = n \log(2\pi) + \log |\tilde{V}| + (y - X\hat{\beta})' \tilde{V}^{-1} (y - X\hat{\beta}). \quad (12)$$

A expressão (12) é a função objectivo para a ML utilizada pelo "proc mixed" do SAS. Minimizando-se essa função sobre todos os parâmetros desconhecidos, obtém-se um sistema de equações cuja solução fornece as estimativas ML. Essas equações são não lineares e são resolvidas numericamente, em geral por processos iterativos como o algoritmo de Newton-Raphson. O processo é repetido até que o critério de convergência adoptado seja satisfeito. Assim, o método da ML supõe normalidade dos dados, é iterativo e fornece sempre estimativas não negativas de componentes de variância, mas estas são enviesadas, pois o método não considera a perda de graus de liberdade resultante da estimação dos efeitos fixos do modelo.

Método da Máxima Verosimilhança Restrita (REML)

Patterson e Thompson (1971) propuseram uma modificação do método da ML para MLGM. Os estimadores REML maximizam a função de verosimilhança de um vector de combinações lineares das observações que são invariantes a $X\beta$. Seja $K'y$ esse vector. Então $K'y = K'X\beta + K'Zu + K'\epsilon$ é invariante a $X\beta$ se e somente se $K'X = \phi$. Com $y \sim N(X\beta, V)$, tem-se que para $K'X = \phi$, $K'y \sim N(\phi, K'VK)$. As equações REML também podem ser deduzidas das equações ML substituindo-se: y por $K'y$, X por $K'X=0$, Z por $K'Z$ e V por $K'VK$. O procedimento "proc mixed" do SAS implementa o método REML construindo a função $-2 \log$ da função de verosimilhança restrita, dada por:

$$-2 \log L_r = (n-k) \log(2\pi) + \log |V| + (y - X\hat{\beta})' V^{-1} (y - X\hat{\beta}) + \log |K'V^{-1}K|, \quad (13)$$

onde k é a característica da matriz X e $\tilde{\beta} = (X' \tilde{V}^{-1} X)^{-1} X' \tilde{V}^{-1} y$, com $\tilde{\beta}$ e \tilde{V} a representar as estimativas REML de β e V , respectivamente.

No método REML a função de verosimilhança é dividida em duas partes independentes, uma referente aos efeitos fixos e outra aos aleatórios, de maneira que a função de verosimilhança é dada pela soma das funções de

verosimilhança de cada parte.

O método REML tem sido considerado o método preferido para estimar componentes de variância de dados desbalanceados (Harville, 1977; Henderson, 1984; Searle et al., 1992, entre outros). As razões para essa preferência são justificadas pelas propriedades desses estimadores. O método REML supõe a normalidade dos dados, é iterativo e fornece sempre estimativas não negativas das componentes de variância, como o método ML. No entanto, considera a perda de graus de liberdade devido aos efeitos fixos, fornecendo estimadores não enviesados e de variância mínima para dados balanceados. A principal diferença entre os métodos ML e REML é que o ML usa a função de verosimilhança de $K'y$ ou o logaritmo desta função, enquanto o REML adopta a função de verosimilhança de y , o qual é um vector de combinações lineares das observações (com média nula), que representa efectivamente as observações ajustadas para os efeitos fixos.

Método de Estimação Quadrática Não Enviesada de Variância Mínima (MIVQUE)

Rao (1971; 1972) propôs um método de estimação que é derivado de modo que o estimador seja uma forma quadrática das observações, não enviesado e de variância mínima. O seu desenvolvimento envolve álgebra extensiva e o seu conceito utiliza valores escolhidos, a priori, para as componentes de variância a estimar. Assim, diferentes valores a priori podem levar a diferentes estimativas para um mesmo conjunto de dados. Obtém-se portanto "um" estimador MIVQUE e não "o" estimador MIVQUE.

Swallow e Monahan (1984) utilizaram o procedimento MIVQUE com a hipótese a priori de que a matriz de variâncias e covariâncias é a matriz identidade, MIVQUEO.

Sob normalidade, a estimação das componentes de variância pelo método MIVQUEO é feita com base na

FACTOR A GÊNERO	FACTOR B = LOCAL DE RESIDÊNCIA		
	URBANO (J=1)	SEMIURBANO (J=2)	RURAL (J=3)
MASCULINO (I=1)	176	153	207
	229	173	177
	197	189	185
	212	195	220
	190	-	-
FEMININO (I=2)	235	220	230
	266	252	249
	280	-	199
	293	-	-
	-	-	-

tabela 1: Valores dos níveis de colesterol de 22 pacientes com diagnóstico clínico de acidente vascular cerebral isquêmico permanente

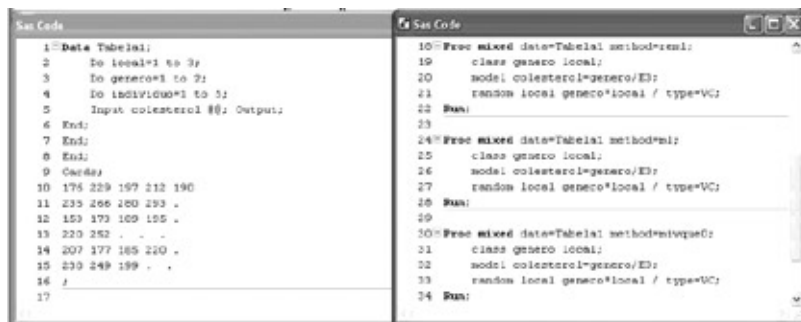


fig 1: Código SAS necessário para ajustar o modelo (15) aos dados da tabela 1

equação:

$$\{tr(PV_i PV_i)\} \hat{S} = \{\hat{S} P V_i P y_i\}, \quad (14)$$

onde \hat{V} é o vector de soluções das componentes de variância e $P = I - X(X'X)^{-1}X'$.

Estruturas gerais de covariância

Conforme referido acima, a análise de modelos mistos envolve duas partes: a análise da parte fixa e da parte aleatória. Tanto a estimação dos efeitos fixos como a predição dos efeitos aleatórios depende da estimação das componentes de variância. A estimação dessas componentes depende da estrutura da matriz G e do método de estimação utilizado. Várias estruturas de covariância podem ser especificadas para a matriz G (a este respeito ver Searle et al., 1992; Wolfinger, 1993, 1996; Littell et al., 2006; entre outros).

Neste estudo, decidiu-se adotar $R = \sigma^2 I$ e as seguintes estruturas de G: componentes de variância (VC), diagonal (TOEP(1)), simetria composta (diagonal mais covariância comum: CS) e Huynh-Feldt (HF).

Procedimento “proc mixed” do SAS

O “proc mixed” é o procedimento do SAS apropriado para a análise de modelos mistos desbalanceados, pois distingue claramente os efeitos fixos dos efeitos aleatórios (Littell et al., 2006). Este procedimento utilizado para ajuste de MLGM permite uma especificação geral da matriz de variâncias e covariâncias e ajusta o MLGM através do método dos MMQG.

O “proc mixed” ajusta o MLGM (1) com a flexibilidade de modelar não somente as médias dos dados, mas tam-

bém as suas variâncias e covariâncias. Para além disso, este procedimento permite uma especificação geral da matriz de variâncias e covariâncias dos erros e que as componentes de erro sejam correlacionados e oferece, ainda, várias opções para a estrutura de variâncias e covariâncias dos efeitos aleatórios, sendo que essas podem ser estimadas, por exemplo, através dos métodos MIVQUEO, ML e REML.

ESTUDO EMPÍRICO

Em seguida, apresenta-se uma aplicação prática com o objectivo de ilustrar a utilização do procedimento “proc mixed” no ajuste de modelos mistos desbalanceados com dois factores (A fixo e B aleatório) e interacção. Os dados apresentados na tabela 1, referem-se aos valores dos níveis de colesterol de 22 pacientes com diagnóstico clínico de acidente vascular cerebral isquêmico permanente, com idade superior ou igual a 60 anos, observados no Hospital Distrital de Faro e classificados de acordo com o género e o local de residência.

O modelo utilizado para descrever os dados da tabela 1 é o modelo misto desbalanceado com dois factores (género fixo e local aleatório) e interacção, caracterizado por:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + Y_{ij} + \epsilon_{ijk}, \quad \text{com } i=1, 2; j=1, 2, 3 \text{ e } k=1, 2, \dots, n_{ij} \quad (15)$$

onde Y_{ijk} é o nível de colesterol do k-ésimo indivíduo do j-ésimo local e do género i; μ é a média global; α_i é o efeito do género; β_j é o efeito do local com $\beta_j \sim N(0; \sigma_\beta^2)$ e independentes; Y_{ij} é o efeito cruzado do local e género, com $Y_{ij} \sim N(0; \sigma_y^2)$ e independentes; ϵ_{ijk} é o erro aleatório, com $\epsilon_{ijk} \sim N(0; \sigma^2)$ e independentes. Admite-se também que β_j, Y_{ij} e ϵ_{ijk} são independentes. Assim:

$$y \sim N(X\beta, ZGZ' + \sigma^2 I) \quad \text{onde } \begin{bmatrix} \sigma_\beta^2 I_j & \phi \\ \phi & \sigma_y^2 I_G \end{bmatrix}$$

Na figura 1 são apresentados os programas utilizando o “proc mixed” do SAS para ajustar os dados da tabela 1. Observa-se que os efeitos fixos são especificados no “model”, e os

aleatórios no "random". A estrutura de G é definida no "random", por meio da opção "type".

No "proc mixed" o método REML é o método de estimação utilizado por

dimento é VC. As outras estruturas de G consideradas neste trabalho são: CS, TOEP(1) e HF.

Com o "proc mixed" estimam-se as componentes de variância para os

para os efeitos fixos. Na tabela 2 são apresentadas algumas informações sobre o modelo misto ajustado pelo "proc mixed" do SAS.

Na figura 2 são apresentados os resultados obtidos por meio do "proc mixed" utilizando a estrutura de componentes de variância para G e o método REML para estimação das componentes de variância. Observa-se que o processo numérico para obter as estimativas REML convergiu na terceira iteração, fornecendo as estimativas das componentes de variância.

Assim, as estimativas pelo método REML, quando G=VC são as seguintes: $\hat{\sigma}^2_{local}=118,05$, $\hat{\sigma}^2_{g\u00e9nero*local}=80,93$ e $\hat{\sigma}^2=469,81$. Portanto, a estimativa de G é dada por:

$$\begin{bmatrix} 118,05I_3 & \phi \\ \phi & 80,93I_8 \end{bmatrix}$$

As funções estimáveis dos tipos I e III são iguais (lemma, 1997), e assim, a hipótese que está sendo testada para o factor fixo, sexo, é dada por $H_0: \alpha_1 - \alpha_2 = 0$. Observa-se na figura 2, que foram estimados três componentes de variância. Os resultados obtidos por meio do "proc mixed" utilizando as outras estruturas de G e os outros métodos são análogos aos da figura 2 e não são apresentados neste artigo.

Segundo Wolfinger (1993) um dos procedimentos utilizado na selecção da estrutura de variâncias e covariâncias é o critério AIC, no qual menores valores sugerem uma estrutura melhor. Os resultados do critério AIC e dos testes de hipóteses para o factor de efeitos fixos, o género, de acordo com as estruturas adoptadas para a matriz G e os métodos de estimação utilizados, estão resumidos na tabela 3.

Quando se comparam as estruturas de G, observa-se que para os três métodos a "melhor", no sentido de menor valor AIC, é a TOEP(1) e a "pioir" é a estrutura HF. No entanto, a escolha da estrutura de G mais apropriada, não deve ser exclusivamente baseada nestes critérios, devem-se também e principalmente considerar a natureza dos dados e a experiência do investigador.

DESCRIÇÃO	INTERPRETAÇÃO
-2 Log verosimilhança restrita	-2 vezes o valor máximo do logaritmo neperiano da verosimilhança restrita (-2l).
Critério de Informação de Akaike (AIC)	O critério AIC pode ser usado para comparar modelos com os mesmos efeitos fixos, mas diferentes estruturas de variância. O modelo com menor AIC é considerado o melhor. $AIC = -2(l - q)$, onde q é o número de parâmetros de covariância existentes em G e R.
Critério de Informação Consistente de Akaike (AICC)	O modelo com menor AICC é considerado o melhor. $AICC = -2(l + q \ln(n^*) + 1)$, onde $n^* = n$ para ML e $n^* = n - p$ para REML, q é o número de parâmetros de covariância e p a ordem de X.
Critério Bayesiano de Schwarz (BIC)	O modelo com menor BIC é considerado o melhor. $BIC = -2(l + q \ln(n^*))$, onde $n^* = n$ para ML e $n^* = n - p$ para REML, q é o número de parâmetros de covariância e p a ordem de X.

tabela 2: Algumas informações sobre o ajuste do MLGM fornecidas pelo "proc mixed"

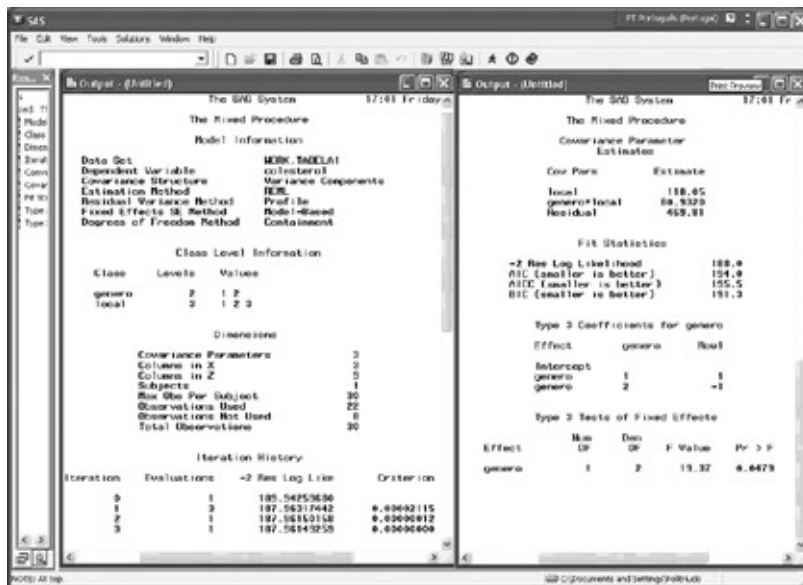


fig 2: Resultados obtidos no SAS quando se ajusta o modelo (15) utilizando o método REML, a opção E3 e G=VC

defeito para se estimar as componentes de variância. No "proc mixed" é possível especificar várias estruturas para a matriz de variâncias e covariâncias dos efeitos aleatórios. A estrutura de G utilizada por defeito neste proce-

defeito para se estimar as componentes de variância do modelo de acordo com o método especificado, bem como as funções estimáveis dos tipos I e III conforme as opções E e E3 especificadas no "model" e o teste F usando as somas de quadrados dos tipos I e III

MÉTODO	ESTRUTURA DE G	VALOR DO AIC	F TIPO III	PR > F
REML	VC	194,0	19,37	0,0479
	CS	194,0	17,87	0,0517
	TOEP(1)	192,0	17,87	0,0517
	HF	210,0	17,87	0,0517
ML	VC	210,5	30,35	0,0314
	CS	(*)	(*)	(*)
	TOEP(1)	208,6	23,76	0,0396
	HF	(*)	(*)	(*)
MIVQUE0	VC	194,0	22,69	0,0414
	CS	194,0	17,33	0,0531
	TOEP(1)	192,0	17,33	0,0531
	HF	218,5	1,00	0,4229

(*) O Processo numérico para obter as estimativas ML não convergiu.

tabela 3: Resultados do AIC e dos testes para o factor de efeitos fixos, obtidos pelo "proc mixed"

O critério AIC não é adequado para comparar os métodos: MIVQUE0, ML e REML, e, a menos que se conheça a matriz de variâncias e covariâncias da população, torna-se pouco confortável para o investigador afirmar qual é o melhor método. Uma comparação analítica torna-se viável para dados balanceados e mesmo assim deve ser feita com cautela, considerando-se cada modelo em particular. Conforme visto anteriormente, o método REML tem sido considerado o preferido para estimar componentes de variância de

dados desbalanceados.

Os testes de hipóteses sobre os efeitos fixos dependem da estrutura da matriz G e do método de estimação utilizado. Comparando-se os testes F, ao nível de significância de 5%, de acordo com as estruturas de G, para cada método separadamente, tem-se: para o MIVQUE0, o teste F foi significativo para G=VC, sendo a estrutura G=HF a mais discrepante de todas; para o ML, o teste F foi significativo para as estruturas de G iguais a VC e TOEP(1); para o

REML, o teste F foi significativo para a estrutura de G=VC, e não significativo para as demais estruturas.

CONCLUSÕES

O SAS apresenta grande flexibilidade no ajuste de MLGM, destacando-se a excelente performance do procedimento "proc mixed".

O "proc mixed" é o procedimento do SAS apropriado para análise de modelos mistos, pois permite uma especificação geral da matriz de variâncias e covariâncias dos efeitos aleatórios, G, e dos resíduos. A estrutura de G é definida na opção "type" da declaração "random". A "melhor" estrutura de G, de entre as utilizadas, no sentido de maior valor AIC, foi G=TOEP(1), para os três métodos de estimação: MIVQUE0, ML e REML.

O nível nominal dos testes de hipóteses do tipo III para o factor de efeitos fixos, género, foram semelhantes quando G=VC. No que se refere às demais estruturas utilizadas neste estudo, apenas a estrutura G=HF alterou sensivelmente o nível nominal, em relação às demais.

1 Investigadora do Centro de Estudos e Investigação em Saúde da Universidade de Coimbra (CEISUC)

REFERÊNCIAS BIBLIOGRÁFICAS

- GILMOUR, A.R.; THOMPSON, R. e CULLIS, B.R. (1995), «Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models», in *Biometrics*, vol.51, 1440-1450.
- FULLER, W.A., e BATTESE, G.E. (1973), «Transformations for estimation of linear models with nested-error structure», in *Journal of the American Statistical Association*, vol. 68, 626-632.
- HARTLEY, H.O. e RAO, J.N.K. (1967), «Maximum likelihood estimation for the mixed analysis of variance model», in *Biometrika*, vol. 54, 93-108.
- HARVILLE, D.A. (1977), «Maximum-likelihood approaches to variances component estimation and to related problems», in *Journal of the American Statistical Association*, vol.72, 320-340.

- HARVILLE, D.A. (1990), «BLUP (Best linear unbiased prediction), and beyond», in *Advances in statistical methods for genetic improvement of livestock*, Springer-Verlag, 239-276.
- HENDERSON, C.R. (1953), «Estimation of variance and covariance components», in *Biometrics*, vol. 9, 226-252.
- HENDERSON, C.R. (1984), Applications of linear models in animal breeding, University of Guelph, Ontario.
- IEMMA, A.F. (1997), Hypothèses statistiques et fonctions estimables du logiciel SAS, Gembloux.
- LITTELL, R.C.; MILLIKEN, G.A.; STROUP, W.W.; WOLFINGER, R.D. e SCHABENBERGER, O. (2006), SAS for mixed models, SAS Institute Inc., Cary.
- MCLEAN, R.A.; SANDERS, W.L. e STROUP, W.W. (1991), «A unified approach to mixed linear models», in *The American Statistician*, vol. 45, 54-64.
- PATTERSON, H.D. e THOMPSON, R. (1971), «Recovery of inter-block information when blocks sizes are unequal», in *Biometrika*, vol. 58, 545-554.
- RAO, C.R. (1971), «Minimum variance quadratic unbiased estimation of variance components», in *Journal of Multivariate Analysis*, vol. 1, 445-456.
- RAO, C.R. (1972), «Estimation of variance and covariance components in linear models», in *Journal of the American Statistical Association*, vol. 67, 112-115.
- RICHARDSON, A.M. e WELSH, A.H. (1995), «Robust restricted maximum likelihood in mixed linear models», in *Biometrics*, vol. 51, 1429-1439.
- ROBINSON, G.K. (1991), «That BLUP is a good thing: the estimation of random effects», in *Statistical Science*, vol. 1, 15-51.
- SCHEFFÉ, H. (1959), The analysis of variance, John Wiley & Sons, New York.
- SEARLE, S.R. (1971), Linear models, John Wiley & Sons, New York.
- SEARLE, S.R. (1987), Linear models for unbalanced data, John Wiley & Sons, New York.
- SEARLE, S.R.; CASELLA, G. e McCULLOCH, C. (1992), Variance components, John Wiley & Sons, New York.
- SWALLOW, W.H. e MONAHAN, J.F. (1984), «Monte Carlo comparison of ANOVA, MINQUE, REML, and ML estimators of variance components», in *Technometrics*, vol. 26, 47-57.
- WOLFINGER, R.D. (1993), «Covariance structure selection in general mixed models», in *Communications in statistics simulation and computation*, vol. 22, n.º 4: 1079-1106.
- WOLFINGER, R.D. (1996), «Heterogeneous variance-covariance structures for repeated measures», in *Journal of Agricultural, Biological and Environmental Statistics*, vol. 1, n.º 2, 205-230.