HIPÓTESES ESTATÍSTICAS COM DADOS DESBALANCEADOS NOS MODELOS COM TRÊS FATORES DE EFEITOS FIXOS HIERARQUIZADOS[1]

Sérgio Minoru Oikawa[2]; Antonio Francisco Iemma[3,4]*

[2]Depto. de Matemática - FCT/UNESP, C.P. 957 - CEP: 19060-900 - Presidente Prudente, SP.
[3]Depto. de Ciências Exatas - ESALQ/USP, C.P. 9 - CEP:13418-900 - Piracicaba, SP
[4]ASSER - Centro de Ensino Superior de São Carlos, SP.
**e-mail: anfiemma@esqualo.esalq.usp.br*

RESUMO: Este trabalho tem por objetivo formalizar os termos das respectivas somas de quadrados e hipóteses mais usuais, que são testadas nos modelos com três fatores de efeitos fixos hierarquizados para dados desbalanceados. Discute-se, também, o problema da interpretação de hipóteses associadas às somas de quadrados, bem como comparam-se os resultados fornecidos por alguns *softwares* estatísticos.
Palavras-chave: dados desbalanceados, somas de quadrados, hipóteses testáveis, *softwares* estatísticos

STATISTICAL HYPOTHESES WITH UNBALANCED DATA IN MODELS WITH THREE FACTORS OF FIXED NESTED EFFECTS

ABSTRATC: The aim of this work is to formalize the terms of the respective sums of squares and more usual hypotheses tested in the models with three factors of fixed nested effects, for unbalanced data. It also discusses the problem of the interpretation of the hypotheses associated to the sums of squares, and comparisons are made for the results provided by some statistical softwares.
Key words: unbalanced data, sums of squares, testable hypotheses, statistical *softwares*

INTRODUÇÃO

Os *softwares* estatísticos tornaram-se uma ferramenta importante e indispensável na análise estatística de dados, principalmente, devido à capacidade dos computadores de hoje, tais como grande rapidez, baixo custo operacional por unidade aritmética e facilidade de acesso. Essas características marcantes fizeram com que o número de usuários de *softwares* estatísticos crescesse consideravelmente. Infelizmente, as dificuldades encontradas por tais usuários têm sido agravadas pela falta de informações detalhadas sobre interpretações de hipóteses nos modelos mais complexos com dados desbalanceados, tanto na literatura quanto nos manuais de utilização.

Nesse contexto, visando amenizar esse problema, um dos objetivos naturais deste trabalho é o estudo dos modelos com três fatores de efeitos fixos hierarquizados para dados desbalanceados. Com base nesse modelo, formalizam-se os termos das respectivas somas de quadrados e as hipóteses mais usuais que são testadas para os efeitos principais. Discute-se, também, o problema da interpretação de hipóteses associadas às somas de quadrados fornecidas por alguns sistemas computacionais estatísticos universalmente consagrados, bem como suas performances em relação ao tema deste estudo.

REVISÃO DE LITERATURA

De acordo com Herr (1986), as análises para ensaios fatoriais com dados desbalanceados, iniciaram-se com as publicações de Yates (1933 e 1934) que, sem dúvida, são um marco no estudo de estimação e testes de hipóteses. Segundo o autor, a maioria dos métodos hoje utilizados são derivados desses dois artigos. Yates (1934) propôs e discutiu quatro métodos para analisar modelos com classificações duplas cruzadas (Two-Way):

$$y_{rst} = \mu + \alpha_r + \beta_s + (\alpha\beta)_{rs} + \varepsilon_{rst},$$

$$r = 1, 2, \ldots, p \; ; \; s = 1, 2, \ldots, q \; ; \; t = 1, 2, \ldots, n_{rs}.$$

[1]Parte da Tese de Doutorado do primeiro autor apresentada à ESALQ/USP - Piracicaba, SP.

Segundo Iemma (1993, 1995/a e 1995b), em 1976 o PROC GLM do SAS incorporou três desses métodos. São eles:

• Método para freqüências das classes proporcionais (Y_1): fornece somas de quadrados não ajustadas para demais fatores. Nesse caso, as somas de quadrados apropriadas para testar efeitos principais podem ser calculadas pelo método descrito para classificações simples (One-Way). As somas de quadrados obtidas através do método Y_1 são equivalentes às somas de quadrados do tipo I, R ($\alpha \mid \mu$) e R ($\beta \mid \mu$), fornecida pelo SAS-GLM. Testam, portanto, as hipóteses do tipo I sobre as médias ponderadas não ajustadas.

• Método do ajustamento de constantes (Y_2): fornece somas de quadrados ajustadas para todos os fatores e interações, exceto interações e/ou fatores hierarquizados que envolvem o fator de interesse. Suas somas de quadrados eqüivalem às somas de quadrados do tipo II, R ($\alpha \mid \mu, \beta$) e R ($\beta \mid \mu, \alpha$), fornecida pelo SAS-GLM e testam as hipóteses do tipo II sobre as médias ponderadas ajustadas.

• Método dos quadrados de médias ponderadas (Y_3): fornece somas de quadrados ajustadas para todos os efeitos envolvidos no modelo com restrição paramétrica do tipo sigma (Modelo-Σ). Corresponde às somas de quadrados do tipo III, R [$\alpha^* \mid \mu^*, \beta^*, (\alpha\beta)^*$] e R [$\beta^* \mid \mu^*, \alpha^*, (\alpha\beta)^*$], fornecida pelo SAS-GLM e testam as hipóteses do tipo III sobre as médias não ponderadas.

Na década de 70, em virtude das confusões na interpretação de hipóteses nos experimentos com dados desbalanceados, bem como à existência de *softwares* que fornecem diferentes resultados para o mesmo conjunto de dados, foram publicados vários artigos sobre ensaios fatoriais.

Francis (1973) constata que quando as análises de variâncias, com dados desbalanceados, são feitas através de diferentes *softwares* estatísticos, os resultados obtidos para as somas de quadrados, além de não serem os mesmos, algumas vezes, são incorretos. Toma como exemplo, um modelo com dois fatores cruzados (A, B) e interação.

Elliott & Woodward (1986) comparam cinco *softwares* estatísticos, para os quais avaliam as respectivas somas de quadrados e hipóteses testadas sob várias opções dos programas para modelos com dois fatores cruzados e interação.

Iemma (1993 e 1995/a) apresenta as hipóteses mais comuns para testar os efeitos de linhas, colunas e interação no modelo com dois fatores cruzados em presença ou não de caselas vazias. Ademais, discute o comportamento do PROC GLM do SAS, versão 6.04, em relação aos métodos e hipóteses testadas.

Santos (1994) estabelece a estruturação das hipóteses e somas de quadrados a elas associadas nos modelos com dois fatores cruzados para dados desbalanceados e apresenta uma revisão dos métodos de análise. Finalizando, o autor compara as saídas geradas pelos *softwares* mais utilizados nas ciências agrárias.

Dallal (1992) apresenta um exemplo mostrando algumas dificuldades na interpretação das somas de quadrados produzidas pelos *softwares* estatísticos, quando se consideram modelos mais complexos. É sem dúvida um dos primeiros trabalhos que envolvem fatores com estruturas cruzadas e hierárquicas para dados desbalanceados. Relatou tais dificuldades, analisando os dados através de dois *softwares*, SAS-GLM, versão 6.04 e SPSS-MANOVA, versão 4.0 para dois tipos de modelos estatisticamente similares,

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\varepsilon_{ijk} \qquad \text{(Modelo-S)}$$

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\gamma_{k(ij)} \qquad \text{(Modelo-C)}$$

onde, o Modelo-S é o modelo tradicional com dois fatores cruzados e interação A*B. O Modelo-C é aquele no qual ele introduz um terceiro fator para classificar as observações dentro de cada casela (i, j) como fator C aninhado sob a interação A*B. Nesse contexto, constata que as somas de quadrados do tipo III, referentes ao fator B fornecidas pelo SAS-GLM, não são as mesmas nos dois modelos. Segundo o autor, esperava-se, no entanto, que elas fossem iguais, pois à exceção de uma simples mudança de classificação, os dois modelos são estatisticamente similares e, portanto, não deveria ocorrer a diferença.

Searle (1994) discute várias razões para a diferença das somas de quadrados do exemplo de Dallal (1992), quando se usa o SAS-GLM e modelo de efeitos fixos hierarquizados em presença de esquema fatorial. Após analisar outros *softwares*, tais como, o BMDP, o SPSS, o SYSTAT e o STATA, o autor conclui que as empresas fabricantes de *softwares* deveriam fornecer através dos manuais, claramente, mais detalhes e, especialmente, mais descrições específicas sobre o que seus *softwares* estão calculando.

METODOLOGIA

Modelos com três fatores hierarquizados

Visando evitar generalizações complexas, apresenta-se aqui o modelo, com base num experimento genérico com três fatores A, B e C contendo a, b_i e c_{ij} níveis, respectivamente. Diz-se que, a estrutura é do tipo estritamente hierárquica quando os c_{ij} níveis do fator C estão aninhados sob cada nível do fator B, sendo que os b_i níveis do fator B estão aninhados sob cada nível do fator A, conforme pode ser visto no esquema da Figura 1.

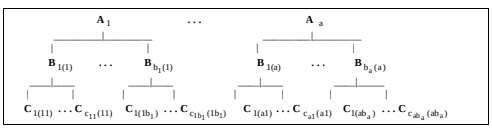

Figura 1 - Esquema para modelos com estrutura do tipo estritamente hierárquica.

Modelos de médias de caselas (Modelo-M)

Segundo Iemma (1997), o modelo de médias de caselas simplifica a construção e a interpretação das hipóteses testadas. Sendo assim, para as classificações hierárquicas com três fatores, o modelo de médias pode ser descrito na forma matricial por:

$$y=W\mu+\varepsilon \quad (1)$$

onde, y é um vetor de realizações de variáveis aleatórias de dimensão (n x 1), W é uma matriz conhecida de "uns" e "zeros" de dimensão (n x $c_{\bullet\bullet}$), com $c_{\bullet\bullet}=\sum_{i=1}^{a}\sum_{j=1}^{b_i} c_{ij}$, μ é um vetor de parâmetros das médias populacionais de dimensão ($c_{\bullet\bullet}$ x 1) e ε é um vetor de variáveis aleatórias não observáveis de dimensão (n x 1), tal que $\varepsilon \sim N(\phi, I\sigma^2)$. O modelo descrito em (1), pode ser caracterizado por:

$$y_{ijtkt}=\mu_{ijk}=\varepsilon_{ijkt}(i=1, ..., a;j=1, ..., b_i;k=1, ..., c_{ij}:t=1, ..., n_{ijk}) \quad (2)$$

onde y_{ijkt} é a resposta observada na t-ésima parcela da casela (i, j, k); μ_{ijk} é média populacional da qual foi retirada a amostra que compõe a casela (i, j, k) e ε_{ijkt} é o erro aleatório atribuído à observação y_{ijtkt} tal que $\varepsilon_{ijkt} \sim NIID(0,\sigma^2)$. Definido o modelo em (1), através do método de mínimos quadrados, obtém-se o Sistema de Equações Normais (SEN), W'Wμ=W'y. Como W tem posto completo, o SEN apresenta solução única para o vetor de médias, dada por $\hat{\mu}=(W'W)^{-1}W'y$.

Modelo superparametrizado (Modelo-S)

Segundo Iemma (1995c) e Iemma & Perri (1997), entre outros, o Modelo-S é parte integrante da história dos modelos lineares e têm sido de grande valia para os pesquisadores das ciências aplicadas, pois exibe explicitamente os parâmetros sobre os quais concentram-se as hipóteses de interesse. Seguindo a caracterização adotada por Winer (1971) entre outros, tem-se:

$$y_{ijkt}=\mu+\alpha_i+\beta_{j(i)}+\gamma_{k(ij)}+\varepsilon_{ijkt} \quad (3)$$

$$i=1, ..., a;j=1, ..., b_i;k=1, ..., c_{ij};t=1, ..., n_{ijk}$$

onde y_{ijkt} e ε_{ijkt}, são como definidos em (2), α_i é o efeito devido ao i-ésimo nível do fator A, $\beta_{j(i)}$ é o efeito devido ao j-ésimo nível do fator B aninhado sob o i-ésimo nível do fator A e $\gamma_{k(ij)}$ é o efeito devido ao k-ésimo nível do fator C aninhado sob o j-ésimo nível do fator B. Descrevendo o modelo (3) na forma matricial, y=Xθ+ε, o SEN, X'Xθ=X'y, a menos de reparametrizações, é indeterminado, pois X é de posto incompleto (Iemma, 1987). Uma solução dentre outras é dada por, $\theta°=(X'X)^{-}$ X'y, onde $(X'X)^{-}$, é uma inversa generalizada qualquer de X'X.

Uma alternativa interessante consiste em utilizar as restrições não estimáveis do tipo $\sum_i \alpha_i = \sum_j \beta_{j(i)} = \sum_k \gamma_{k(ij)} = 0$ a fim de reparametrizar o modelo (3) obtendo-se o modelo reparametrizado de posto completo (Modelo-$\sum$),

$$y_{ijkt}=\mu^*+\alpha_i^*+\beta^*_{j(i)}+\gamma^*_{k(ij)}+\varepsilon_{ijkt} \quad (4)$$

$$i=1, ..., (a-1); j=, ..., (b_i-1);k=1, ..., (c_{ij}-1);t=1, ..., (n_{ijk}-1)$$

De modo análogo ao modelo de médias de caselas, o Sistema de Equações Normais, X*'X*θ* = X*'y, apresenta solução única, $\hat{\theta}^*=(X^{*\prime}X^*)^{-1}X^{*\prime}y$, pois X* tem posto coluna completo. Ademais, se os dados são desbalanceados com todas as caselas ocupadas, o modelo-$\sum$ fornece somas de quadrados, ajustadas para todos os efeitos envolvidos, eqüivalentes ao método dos quadrados de médias ponderadas de Yates.

Obtenção das somas de quadrados

Conforme o interesse, os modelos M, S e $\sum$ podem assumir diversas caracterizações. Por exemplo, no Modelo-S, tem-se;

$$y=X_1\theta_1+\varepsilon \Leftrightarrow y_{ijkt}=\mu+\varepsilon_{ijkt} \quad (S.11)$$

$$y=X_2\theta_2+\varepsilon \Leftrightarrow y_{ijkt}=\mu+\alpha_i+\varepsilon_{ijkt} \quad (S.12)$$

$$y=X_3\theta_3+\varepsilon \Leftrightarrow y_{ijkt}=\mu+\alpha_i+\beta_{j(i)}+\varepsilon_{ijkt} \quad (S.13)$$

$$y=X\theta+\varepsilon \Leftrightarrow y_{ijkt}=\mu+\alpha_i+\beta_{j(i)}+\gamma_{k(ij)}+\varepsilon_{ijkt} \quad (S.14)$$

Tais parametrizações sucessivas e ordenadas facilitam a interpretação da notação-R (.) e de certas somas de quadrados a elas associadas. A notação-R(.) é um procedimento para a obtenção de somas de quadrados, através da redução da soma de quadrados totais ao ajustar um modelo particular de interesse. O termo R(.) é a medida de variação em y explicada pelo modelo ajustado.

Logo, a redução da soma de quadrados, por exemplo, devida ao ajuste do modelo (S.14), sugerida por Searle (1987) como notação-R(.) é dada por:

$$R(\theta) = y'X(X'X)^{-} X'y = \theta^{o\prime}X'y = \text{S.Q.Parâmetros} \quad (5)$$

Sendo assim, a notação R(.) fornece uma medida conveniente para descrever os procedimentos computacionais usados na obtenção das somas de quadrados. Nesse caso,

$$y = X_1\theta_1+\varepsilon \Rightarrow X'_1X_1\theta_1= X'_1y \Rightarrow R(\mu)=\theta_1^{0\prime}X_1'y \quad (S.21)$$

$$y = X_2\theta_2+\varepsilon \Rightarrow X'_2X_2\theta_2=X'_2y \Rightarrow R(\mu,\alpha)=\theta_2^{0\prime}X_2'y \quad (S.22)$$

$$y = X_3\theta_3+\varepsilon \Rightarrow X'_3X_3\theta_3=X'_3y \Rightarrow R[\mu,\alpha.\beta(\alpha)]=\theta_3^{0\prime}X_3'y \quad (S.23)$$

$$y = X\theta+\varepsilon \Rightarrow X'X\theta \Rightarrow X'y \Rightarrow R[\mu,\alpha.\beta(\alpha),\gamma(\alpha)\beta]=\theta^{\circ\prime}X'y \quad (S.24)$$

Procedendo-se os ajustes do tipo seqüencial que o SAS-GLM denota em sua saída por tipo I, têm-se as seguintes somas de quadrados:

$$R(\alpha \mid \mu) = R(\mu, \alpha) - R(\mu) \quad (6)$$

$$R[\beta(\alpha) \mid \mu, \alpha] = R[\mu, \alpha, \beta(\alpha)] - R(\mu, \alpha) \quad (7)$$

$$R[\gamma(\alpha\beta) \mid \mu, \alpha, \beta(\alpha)] = R[\mu, \alpha, \beta(\alpha), \gamma(\alpha\beta)] - R[\mu, \alpha, \beta(\alpha)] \quad (8)$$

Ademais, estendendo-se os procedimentos usuais dados em Searle (1987) para modelos hierarquizados com dois fatores, obtém-se:

$$SQA=\sum_{i=1}^{a} n_{i\bullet\bullet}(\bar{y}_{i\bullet\bullet\bullet}-\bar{y}_{\bullet\bullet\bullet\bullet})^2=R(\alpha|\mu) \quad (9)$$

$$SQB(A)=\sum_{i=1}^{a}\sum_{j=1}^{b_i} n_{ij\bullet}(\bar{y}_{ij\bullet\bullet}-\bar{y}_{i\bullet\bullet\bullet})^2 = R[\beta(\alpha) \mid \mu, \alpha] \quad (10)$$

$$SQC(AB)=\sum_{i=1}^{a}\sum_{j=1}^{b_i}\sum_{k=1}^{c_{ij}} n_{ijk}(\bar{y}_{ijk\bullet}-\bar{y}_{ij\bullet\bullet})^2=R[\gamma(\alpha\beta) \mid \mu, \beta(\alpha)] \quad (11)$$

Infelizmente, as dificuldades encontradas nas interpretações das hipóteses testadas pelas somas de quadrados, utilizando-se tanto a notação-R (.) quanto os procedimentos usuais, é que eles não especificam as hipóteses testadas.

Searle (1987), Iemma (1987) e Iemma & Perri (1997), entre outros, adotam um procedimento alternativo para obter as somas de quadrados baseadas em hipóteses de interesse. Assim, para testar a hipótese do tipo H_o :B'μ=ϕ, onde B'μ é um conjunto de funções estimáveis, B' tem posto linha completo e ϕ é um vetor de "zeros", obtém-se a estatística de Wald:

$$SQH_O=(\mathbf{B}'\hat{\mu})'[\mathbf{B}'(\mathbf{W}'\mathbf{W})^{-1}\mathbf{B}]^{-1}(\mathbf{B}'\hat{\mu}) \quad (12)$$

Tal procedimento é bastante simples quando se adota o modelo de médias de caselas, pois esse modelo facilita a especificação das hipóteses testadas.

Os quatro tipos de funções estimáveis

Conforme Iemma (1993), se a matriz X não tem posto coluna completo, então o vetor θ° não é estimador não viesado no conceito de Rao (1945). De fato, nessas condições $E[\theta^{o}] = (X'X)^{g}X'X\theta=H\theta$. Então, a função $H\theta$ é estimável.

Utilizando essas idéias, o PROC GLM fornece uma base para o estudo da estimabilidade dessas funções. Para tanto, toma $L = (X'X)^{g2}X'X$, analogamente à matriz H, com a diferença que G é definida para uma inversa generalizada qualquer de X'X, não sendo única. Já a inversa generalizada g2 fornece um "L" único. Assim, enquanto no primeiro caso tem-se vários conjuntos equivalentes de funções estimáveis, no segundo tem-se um único que, sem dúvida é um deles.

Com base em Lθ, o PROC GLM gera as funções estimáveis dos tipos I, II, III e IV para obter as hipóteses testáveis dos tipos I, II, III e IV.

Face aos objetivos deste trabalho não será discutido aqui o procedimento de obtenção das inversas e, nem mesmo, as regras para obtenção das funções estimáveis. O tema está discutido com riqueza de detalhes em Mondardo & Iemma (1998).

Hipóteses estatísticas usualmente testadas pelo SAS-GLM

Hipóteses sobre o fator A

Embora, a soma de quadrados, R ($\alpha \mid \mu$), bem como aquela obtida através do procedimento usual, SQA, não especifiquem as hipóteses testadas, elas correspondem à hipótese testada pela soma de quadrados do tipo I fornecida pelo SAS-GLM. Nesse caso, testam a hipótese do tipo I sobre as médias ponderadas não ajustadas. Sua forma geral é:

$$H_0^{(1)}: \sum_{j=1}^{b_i}\sum_{k=1}^{c_{ij}} \frac{n_{ijk}\mu_{ijk}}{n_{i\bullet\bullet}} = \sum_{j=1}^{b_{i'}}\sum_{k=1}^{c_{i'j}} \frac{n_{i'jk}\mu_{i'jk}}{n_{i'\bullet\bullet}}, \forall\ i, i'(i \neq i') \quad (13)$$

A soma de quadrados do tipo II sobre o fator A, fornecida pelo SAS-GLM, não testa a hipótese do tipo II sobre as médias ponderadas ajustadas para os fatores B(A) e C(A B) como nos casos de esquemas fatoriais, pois eles estão aninhados sob o fator A. Sendo assim, a soma de quadrados do tipo II testa hipótese equivalente à do tipo I.

Quando o fator C(A B) tem o mesmo número de níveis dentro de cada nível do fator B(A), independentemente dos dados serem desbalanceados ou não, a soma de quadrados do tipo III fornecida pelo SAS-GLM é equivalente àquela obtida através do método dos quadrados de médias ponderadas proposto por Yates (1934). Testa, portanto, a hipótese do tipo III sobre as médias não ponderadas, dada por:

$$H_0^{(2)}: \mu_{i\bullet\bullet} = \mu_{i'\bullet\bullet}, \forall\ i, i'(i \neq i') \quad (14)$$

Se, entretanto, o número de níveis do fator C(A B) é diferente dentro de cada nível do fator B(A), independentemente dos dados serem balanceados ou não, a soma de quadrados do tipo III não testa a hipótese $H_0^{(2)}$ mas sim uma hipótese bastante complexa, $H_0^{(3)}$gerada a partir de funções estimáveis do tipo III:

$$H_0^{(3)}: \frac{\sum_{j=1}^{b_i}\sum_{k=1}^{c_{ij}}[\mu_{ijk}/(c_{ij}+1)]}{\sum_{j=1}^{b_i}[c_{ij}/(c_{ij}+1)]} = \frac{\sum_{j=1}^{b_{i'}}\sum_{k=1}^{c_{i'j}}[\mu_{i'jk}/(c_{i'j}+1)]}{\sum_{j=1}^{b_{i'}}[c_{i'j}/(c_{i'j}+1)]}, \forall\ i, i'(i \neq i') \quad (15)$$

De acordo com Searle (1987), as somas de quadrados dos tipos I, II e III são obtidas através do ajuste de diferentes parametrizações. A soma de quadrados do tipo IV, no entanto, é gerada pela própria rotina do SAS-GLM, baseando-se nas configurações das caselas ocupadas. Nesse caso, dependendo do número e da posição das caselas vazias, o SAS-GLM pode gerar diferentes somas de quadrados do tipo IV e, portanto, testar diferentes hipóteses do tipo IV. Nos modelos com três fatores hierarquizados, porém, a soma de quadrados do tipo IV referente ao fator A testa sempre a hipótese $H_0^{(2)}$.

Hipóteses sobre o fator B aninhado sob o fator A

A soma de quadrados, R [$\beta(\alpha) \mid \mu, \alpha$] = SQB(A), corresponde às hipóteses testadas pelas somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Nesse caso, independentemente dos dados serem desbalanceados ou não e do fator C(A B) ter números de níveis diferentes ou não, elas testam a hipótese do tipo I sobre as médias ponderadas não ajustadas, dada por:

$$H_0^{(4)}: \left\{ \sum_{k=1}^{c_{ij}} \frac{n_{ijk}\mu_{ijk}}{n_{ij\bullet}} = \sum_{k=1}^{c_{ij'}} \frac{n_{ij'k}\mu_{ij'k}}{n_{ij'\bullet}},\ \forall\ j, j'\ (j \neq j') \right\} \text{ dentro de i.} \quad (16)$$

Já as somas de quadrados dos tipos III e IV fornecidas pelo SAS-GLM testam hipóteses equivalentes do tipo III sobre as médias não ponderadas, mesmo com caselas vazias. Sua forma geral é:

$$H_0^{(5)}: \left\{ \mu_{ij\bullet} = \mu_{ij'\bullet}, \forall\ j, j'\ (j \neq j') \right\} \text{dentro de i.} \quad (17)$$

Hipótese sobre o fator C aninhado sob o fator B(A)

A soma de quadrados R [$\gamma(\alpha\ \beta) \mid \mu, \alpha, \beta(\alpha)$] = SQC(A B) testa hipótese equivalente às hipóteses testadas pelas somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM, pois o fator C está aninhado sob o fator B, sendo que o fator B está aninhado sob o fator A. Sua forma é:

$$H_0^{(1)}: \{(\mu_{ijkt} = \mu_{ijkt'}, \forall K, K'(K \neq K') \text{dentro de j}\} \text{dentro de i.} \quad (18)$$

RESULTADOS E DISCUSSÃO

Exemplo: Para ilustrar os procedimentos descritos, utiliza-se um conjunto de dados sobre comprimento de fibras de eucalipto, adaptado de Padovani (1984) e reproduzido na TABELA 1. Nesse experimento, a espécie utilizada foi o *Eucalyptus grandis* Hill ex Maiden, aos três anos de idade, de povoamento pertencente à Champion Papel e Celulose S/A, instalado no Horto Santa Teresinha, no Município de Mogi-Guaçu. Foram tomadas duas árvores com 10,0 cm de diâmetro à altura do peito (DAP) das quais retiraram-se secções transversais (discos), ao

nível do DAP. Nesses discos, consideraram-se os incrementos anuais de crescimento em três posições, denominadas: posição 1, região próxima à medula; posição 3, região próxima à casca; e posição 2, região intermediária, correspondendo, respectivamente, ao 1º, 2º e 3º anos de crescimento. Para cada árvore, foram obtidas três amostras correspondentes às posições 1, 2 e 3, relativas aos anos de crescimento. A partir desse material macerado, foi procedida a montagem de lâminas em geléia de glicerina. Nas lâminas, foram realizadas medições de fibras inteiras, totalmente ao acaso. O comprimento das fibras foi medido em micra, em microscópio com aumento de cem vezes. Os dados foram adaptados para gerar desbalanceamento com caselas vazias. Sendo assim, foram consideradas duas posições na árvore 2 e número deferente de lâminas.

Obtenção das Somas de Quadrados

O modelo definido em (3) assume diversas caracterizações conforme o interesse. Nesse caso, fazendo-se as parametrizações sucessivas e ordenadas, como em (S.11) até (S.14), foram obtidas as somas de quadrados seqüenciais, apresentadas a seguir.

$R(\alpha|\mu)=0,1016466$; $R\ \beta(\alpha) \mid \mu, \alpha]=0,551434$;
$R\ [\gamma(\alpha\beta) \mid \mu, \alpha,\beta(\alpha)]=0,0331818$

Tais resultados podem, também, ser obtidos através do PROC SAS-GLM, utilizando-se o programa 1, conforme pode ser observado na TABELA 2.

Programa 1: Programa SAS-GLM para modelos com Três Fatores Hierarquizados

```
DATA NEST3;
INPUT A B C Y;
CARDS;
1 1 1 0,791
1 1 1 0,749
      ®
2 2 2 0,915
;
```

TABELA 1 - Comprimento de fibras de eucalipto, em micra.

Árvore (A1)						Árvore(A2)				
Posição						Posição				
B1		B2	B3			B1			B2	
Lâmina		Lâmina	Lâmina			Lâmina			Lâmina	
C1	C2	C1	C1	C2	C3	C1	C2	C3	C1	C2
0,791	0,649	1,729	1,162	0,909	0,979	0,944	0,702	0,708	0,968	1,21
0,749	0,791	0,915	0,785	1,062	0,962	0,82	0,655	0,861	0,956	0,915
-	0,761	0,856	-	0,915	1,239	0,743	0,915	0,743	1,033	-
-	-	1,003	-	1,103	1,033	-	-	0,696	-	-

Fonte: Conjunto de dados adaptado de Padovani(1984).

TABELA 2 - Análise de variância fornecida pelo SAS-GLM.

General Linear Model Procedure					
Dependent Variable:Y					
Source	DF	Sum of Squares	Mean Square	F Value	Pr. > F
Model	10	0,68626255	0,06862626	2,02	0,0789
Error	23	0,78092992	0,03395347		
C. Total	33	1,46719247			
Source	DF	Type I e II SS	Type III SS	Type IV SS	
A	1	0,10164659	0,02167112	0,03163605	
B(A)	3	0,55143413	0,52756684	0,52756684	
C(A B)	6	0,03318183	0,03318183	0,03318183	

```
PROC GLM;
CLASS A B C;
MODEL Y = A  B(A) C(A B) / SS1 SS2 SS3 SS4
E1 E2 E3 E4;
RUN;
```

Hipóteses testadas pelo SAS-GLM no modelo com três fatores hierarquizados

Hipóteses sobre o fator A

A soma de quadrados obtida através de R ($\alpha \mid \mu$) = SQA corresponde às hipóteses testadas pelas somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Nesse caso, ambas testam a hipótese do tipo I sobre as médias ponderadas não ajustadas, descrita em (13) e que para os dados da TABELA 1 resulta em:

$$H_0^{(1)}: \frac{2\mu_{111}+3\mu_{112}+4\mu_{121}+2\mu_{131}+4\mu_{132}+4\mu_{133}}{19} = \frac{3\mu_{211}+3\mu_{212}+4\mu_{213}+3\mu_{221}+2\mu_{222}}{15}$$

A Hipótese$H_0^{(1)}$ pode ser escrita na forma$H_0^{(1)}$:B'μ=ϕ, onde

B'=[2/19 3/19 4/19 2/19 4/19 4/19-3/15-3/15-4/15-3/15-2/15]

e $\mu'=[\mu_{111}\ \mu_{112}\ \mu_{121}\ \mu_{131}\ \mu_{132}\ \mu_{133}\ \mu_{211}\ \mu_{212}\ \mu_{213}\ \mu_{221}\ \mu_{222}]$

Sendo assim, a soma de quadrados associada à hipótese $H_0^{(1)}$ pode ser obtida, como em (12), utilizando a estatística de Wald, $SQH_0^{(1)}=(B'\hat{\mu})'[B'(W'W)^{-1}B]^{-1}(B'\hat{\mu})$ onde

$\hat{\mu}$ = [0,77 0,7337 1,1257 0,9735 0,9972 1,0532 0,8357 0,7573 0,7520 0,9857 1,0625] ;

$(W'W)^{-1}$=diag (1/ n_{ijk}) = diag (1/2, 1/3, 1/4, 1/2, 1/4, 1/4, 1/3, 1/3, 1/4, 1/3, 1/2) ;

B' $\hat{\mu}$ =0,1101193; $[B'(W'W)^{-1}B]$=0,1192982 e $[B'(W'W)^{-1}B]^{-1}$=8,382353

Desse modo,

$SQH_0^{(1)}$= 0,1016466 = R ($\alpha \mid \mu$) = SQA.

Substituindo-se μ_{ijk} por $\mu+\alpha_i+\beta_{j(i)}+\gamma_{k(ij)}$ em $H_0^{(1)}$pois o SAS-GLM utiliza o Modelo-S na apresentação das funções estimáveis, conforme amplamente discutido em Mondardo & Iemma (1998), a hipótese associada à soma de quadrados do tipo I resulta em,

$$H_0^{(1)}: \begin{cases} \alpha_1+\frac{1}{19}(5\beta_{1(1)}+4\beta_{2(1)}+10\beta_{3(1)}+2\gamma_{1(11)}+3\gamma_{2(11)}+4\gamma_{1(12)}+2\gamma_{1(13)}+4\gamma_{2(13)}+4\gamma_{3(13)})= \\ =\alpha_2+\frac{1}{15}(10\beta_{1(2)}+5\beta_{2(2)}+3\gamma_{1(21)}+3\gamma_{2(21)}+4\gamma_{3(21)}+3\gamma_{1(22)}+2\gamma_{2(22)}) \end{cases}$$

como pode ser observado na TABELA 3 fornecida pelo SAS-GLM, fazendo-se L2 = 1. Ali pode ser observado que, como há um único grau de liberdade para o fator A, há apenas um valor para os coeficientes L's, no caso dado por L2. Observe, por exemplo, que 5/19 = 0,2632 ; 4/19=0,2105 e assim por diante. Sem dúvida, $H_0^{(1)}$ é uma hipótese difícil de ser interpretada, especialmente para os pesquisadores não iniciados na teoria dos testes de hipóteses estatísticas.

Conforme descrito no capítulo anterior, como o fator C tem números de níveis diferentes dentro de cada nível do fator B, a soma de quadrados do tipo III, para o fator A, fornecida pelo SAS-GLM não testa a hipótese sobre as médias não ponderadas e, portanto, não é equivalente à hipótese $H_0^{(2)}$. Ao contrário, testa uma hipótese não usual, baseada em funções estimáveis complexas do tipo III, $H_0^{(3)}$, dada como em (15) por:

$$H_0^{(3)}: \frac{\frac{(\mu_{111}+\mu_{112})}{(2+1)}+\frac{\mu_{121}}{(1+1)}+\frac{(\mu_{131}+\mu_{132}+\mu_{133})}{(3+1)}}{\frac{2}{3}+\frac{1}{2}+\frac{3}{4}} = \frac{\frac{(\mu_{211}+\mu_{212}+\mu_{213})}{(3+1)}+\frac{(\mu_{221}+\mu_{222})}{(2+1)}}{\frac{3}{4}+\frac{2}{3}}$$

Através da estatística de Wald obtém-se, como consta da TABELA 2,

$SQH_0^{(3)}$=0,0216711.

Fazendo-se, como em $H_0^{(1)}$, as substituições devidas em termos de Modelo-S, a hipótese $H_0^{(3)}$ resulta na forma fornecida pelo SAS-GLM,

$$H_0^{(3)}: \begin{cases} \alpha_1+\frac{1}{23}(8\beta_{1(1)}+6\beta_{2(1)}+9\beta_{3(1)}+4\gamma_{1(11)}+4\gamma_{2(11)}+6\gamma_{1(12)}+3\gamma_{1(13)}+3\gamma_{2(13)}+3\gamma_{3(13)})= \\ =\alpha_2+\frac{1}{17}(9\beta_{1(2)}+8\beta_{2(2)}+3\gamma_{1(21)}+3\gamma_{2(21)}+3\gamma_{3(21)}+4\gamma_{1(22)}+4\gamma_{2(22)}) \end{cases}$$

como se observa na TABELA 3, fazendo-se L2=1, também associada a 1 g.l..

Embora a hipótese $H_0^{(3)}$ tenha sido formulada neste trabalho, mesmo assim, vem confirmar a suposição de Searle (1994) de que parece não haver uma explicação estatística para que a soma de quadrados do tipo III seja calculada dessa forma pelo PROC SAS-GLM, tanto nos manuais de utilização como na literatura estatística. Realmente, é uma hipótese difícil de ser interpretada.

Segundo Hocking (1985), a ocorrência de somas de quadrados do tipo III que testam hipóteses não usuais, derivada de funções complexas, é conseqüência do procedimento computacional utilizado pelo SAS-GLM, não tendo outra explicação. Para o autor, isso ocorre

TABELA 3 - Funções estimáveis fornecidas pelo SAS-GLM.

General Linear Models Procedure							
Estimable Functions for:							
		A			B(A)		C(A B)
Type		I e II	III	IV	I e II	III e IV	I, II, III e IV
Effects		Coef.	Coef.	Coef.	Coef.	Coef.	Coef.
Interc.		0	0	0	0	0	0
A	1	L2	L2	L2	0	0	0
	2	-L2	-L2	-L2	0	0	0
B(A)	11	0,2632*L2	0,3478*L2	0,3333*L2	L4	L4	0
	21	0,2105*L2	0,2609*L2	0,1667*L2	L5	L5	0
	31	0,5263*L2	0,3913*L2	0,5*L2	-L4-L5	-L4-L5	0
	12	-0,6667*L2	-0,5294*L2	-0,6*L2	L7	L7	0
	22	-0,3333*L2	-0,4706*L2	-0,4*L2	-L7	-L7	0
C(A B)	111	0,1053*L2	0,1739*L2	0,1667*L2	0,4*L4	0,5*L4	L9
	211	0,1579*L2	0,1739*L2	0,1667*L2	0,6*L4	0,5*L4	-L9
	112	0,2105*L2	0,2609*L2	0,1667*L2	L5	L5	0
	113	0,1053*L2	0,1304*L2	0,1667*L2	-0,2*(L4+L5)	-0,3333(L4+L5)	L12
	213	0,2105*L2	0,1304*L2	0,1667*L2	-0,4*(L4+L5)	-0,3333(L4+L5)	L13
	313	0,2105*L2	0,1304*L2	0,1667*L2	-0,4*(L4+L5)	-0,3333(L4+L5)	-L12-L13
	121	-0,2*L2	-0,1765*L2	-0,2*L2	0,3*L7	0,3333*L7	L15
	221	-0,2*L2	-0,1765*L2	-0,2*L2	0,3*L7	0,3333*L7	L16
	321	-0,2667*L2	-0,1765*L2	-0,2*L2	0,4*L7	0,3333*L7	-L15-L16
	122	-0,2*L2	-0,2353*L2	-0,2*L2	-0,6*L7	-0.5*L7	L18
	222	-0,1333*L2	-0,2353*L2	-0,2*L2	-0,4*L7	-0.5*L7	-L18

freqüentemente em modelos que envolvem três ou mais fatores de efeitos fixos.

Já a soma de quadrados do tipo IV testa sempre a hipótese $H_0^{(2)}$ sobre as médias não ponderadas, independentemente de ocorrer caselas vazias ou não. Logo, tem-se:

$$H_0^{(2)}: \frac{\mu_{111}+\mu_{112}+\mu_{121}+\mu_{131}+\mu_{132}+\mu_{133}}{6} = \frac{\mu_{211}+\mu_{212}+\mu_{213}+\mu_{221}+\mu_{222}}{5}$$

Sendo assim,

$$SQH_0^{(2)} = 0{,}031636.$$

Procedendo de modo análogo, obtém-se a forma de $H_0^{(2)}$ fornecida pelo SAS-GLM,

$$H_0^{(2)}: \begin{cases} \alpha_1 + \frac{1}{6}(2\beta_{1(1)} + \beta_{2(1)} + 3\beta_{3(1)} + \gamma_{1(11)} + \gamma_{2(11)} + \gamma_{1(12)} + \gamma_{1(13)} + \gamma_{2(13)} + \gamma_{3(13)}) = \\ = \alpha_2 + \frac{1}{5}(3\beta_{1(2)} + 2\beta_{2(2)} + \gamma_{1(21)} + \gamma_{2(21)} + \gamma_{3(21)} + \gamma_{1(22)} + \gamma_{2(22)}) \end{cases}$$

como se observa na TABELA 3, fazendo-se L2=1.

Hipóteses sobre o fator B aninhado sob o fator A

As somas de quadrados R [β(α) | μ, α] e SQB(A), testam hipóteses equivalentes às hipóteses testadas pelas somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Logo, testam a hipótese do tipo I sobre as médias ponderadas não ajustadas. Uma forma é dada por:

$$H_0^{(4)}: \begin{cases} \frac{2\mu_{111}+3\mu_{112}}{5} = \frac{2\mu_{131}+4\mu_{132}+4\mu_{133}}{10} \\ \mu_{121} = \frac{2\mu_{131}+4\mu_{132}+4\mu_{133}}{10} \\ \frac{3\mu_{211}+3\mu_{212}+4\mu_{213}}{10} = \frac{3\mu_{221}+2\mu_{222}}{5} \end{cases}$$

Desse modo,

$$SQH_0^{(4)} = 0{,}5514341 = R\,[\beta(\alpha) \mid \mu, \alpha] = SQB(A).$$

Em termos de Modelo-S, a hipótese $H_0^{(4)}$ resulta na forma dada pelo SAS-GLM,

$$H_0^{(4)}: \begin{cases} \beta_{1(1)} + \frac{1}{5}(2\gamma_{1(11)} + 3\gamma_{2(11)}) = \beta_{3(1)} + \frac{1}{10}(2\gamma_{1(13)} + 4\gamma_{2(13)} + 4\gamma_{3(13)}) \\ \beta_{2(1)} + \gamma_{1(12)} = \beta_{3(1)} + \frac{1}{10}(2\gamma_{1(13)} + 4\gamma_{2(13)} + 4\gamma_{3(13)}) \\ \beta_{1(2)} + \frac{1}{10}(3\gamma_{1(21)} + 3\gamma_{2(21)} + 4\gamma_{3(21)}) = \beta_{2(2)} + \frac{1}{5}(3\gamma_{1(22)} + 2\gamma_{2(22)}) \end{cases}$$

Como se observa na TABELA 3, agora fazendo-se: L4=1 e L5=L7=0 ; L5=1 e L4=L7=0 ; L4=L5=0 e L7=1, pois o fator B está associada a 3 g.l..

Já as somas de quadrados dos tipos III e IV fornecidas pelo SAS-GLM testam, com 3 g.l., a hipótese do tipo III, aqui denotada $H_0^{(5)}$, e dada por:

$$H_0^{(5)}: \begin{cases} \dfrac{\mu_{111}+\mu_{112}}{2} = \dfrac{\mu_{131}+\mu_{132}+\mu_{133}}{3} \\ \mu_{121} = \dfrac{\mu_{131}+\mu_{132}+\mu_{133}}{3} \\ \dfrac{\mu_{211}+\mu_{212}+\mu_{213}}{3} = \dfrac{\mu_{221}+\mu_{222}}{2} \end{cases}$$

Logo,

$$SQH_0^{(5)} = 0{,}5275668.$$

Descrevendo $H_0^{(5)}$ em termos de Modelo-S, resulta na forma dada pelo SAS-GLM,

$$H_0^{(5)}: \begin{cases} \beta_{1(1)} + \dfrac{1}{2}(\gamma_{1(11)} + \gamma_{2(11)}) = \beta_{3(1)} + \dfrac{1}{3}(\gamma_{1(13)} + \gamma_{2(13)} + \gamma_{3(13)}) \\ \beta_{2(1)} + \gamma_{1(12)} = \beta_{3(1)} + \dfrac{1}{3}(\gamma_{1(13)} + \gamma_{2(13)} + \gamma_{3(13)}) \\ \beta_{1(2)} + \dfrac{1}{3}(\gamma_{1(21)} + \gamma_{2(21)} + \gamma_{3(21)}) = \beta_{2(2)} + \dfrac{1}{2}(\gamma_{1(22)} + \gamma_{2(22)}) \end{cases}$$

como se observa na TABELA 3, fazendo-se: L4=1 e L5=L7=0 ; L5=1 e L4=L7=0 ; L4=L5=0 e L7=1.

Hipóteses sobre o fator C aninhado sob o fator B(A)

A soma de quadrados obtida pelo procedimento usual, SQC(A B), bem como aquela obtida através da notação-R (.), R [$\gamma(\alpha\ \beta) \mid \mu, \alpha, \beta(\alpha)$], testam hipóteses equivalentes às hipóteses testadas pelas somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM. Uma forma, com 6 g.l., é dada por:

$$H_o^{(6)}: \begin{cases} \mu_{111} = \mu_{112} \\ \mu_{131} = \mu_{133} \\ \mu_{132} = \mu_{133} \\ \mu_{211} = \mu_{213} \\ \mu_{212} = \mu_{213} \\ \mu_{221} = \mu_{222} \end{cases} \Leftrightarrow \quad H_o^{(6)}: \begin{cases} \gamma_{1(11)} = \gamma_{2(11)} \\ \gamma_{1(13)} = \gamma_{3(13)} \\ \gamma_{2(13)} = \gamma_{3(13)} \\ \gamma_{1(21)} = \gamma_{3(21)} \\ \gamma_{2(21)} = \gamma_{3(21)} \\ \gamma_{1(22)} = \gamma_{2(22)} \end{cases}$$

Sendo assim,

$$SQH_0^{(6)} = 0{,}0331818 = R\ [\gamma(\alpha\ \beta) \mid \mu, \alpha, \beta(\alpha)] = SQC(A\ B)\ .$$

Em termos do Modelo-S, a hipótese $H_0^{(6)}$ pode ser descrita de modo análogo às anteriores. Basta fazer na TABELA 3, L9=1 e os demais iguais a zero; L12=1 e os demais iguais a zero; L13=1 e os demais iguais a zero; L15=1 e os demais iguais a zero; L16=1 e os demais iguais a zero; finalmente, L18=1 e os demais iguais a zero, pois $H_0^{(6)}$ está associada a 6 g.l..

Hipóteses Testadas Por Outros Softwares nos Modelos com Três Fatores Hierarquizados para Dados Desbalanceados

Com o objetivo de elucidar aos usuários, apresenta-se aqui uma comparação, sem o apelo de competição, das hipóteses testadas através do SAS-GLM com aquelas testadas por outros *Softwares* estatísticos. Para tanto, foram utilizados o MINITAB, o NTIA, o STATGRAPHICS, o SAEG, o GLIM, o STATISTICA e o BMDP.

Ressalta-se aqui o fato de que os *softwares* são abordados do ponto de vista do usuário e não do especialista. Nesse contexto são utilizados apenas comandos básicos usuais e não programação mais sofisticada.

MINITAB – Versão 11.0

Programa 2: Programa MINITAB para modelos com Três Fatores Hierarquizados

```
MTB> NAME C1='A' C2='B' C3='C' C4='Y'
MTB> GLM Y = A B(A) C(A B)
```

Em geral, quando os dados são desbalanceados com todas as caselas ocupadas, o procedimento GLM do MINITAB fornece somas de quadrados dos tipos seqüenciais e ajustadas, equivalentes às somas de quadrados dos tipos I e III fornecidas pelo SAS-GLM. Entretanto, para os dados da TABELA 1, o procedimento GLM fornece apenas as somas de quadrados seqüenciais e não realiza nenhum teste estatístico, pois os níveis do fator C aninhado sob o fator B não são os mesmos dentro de cada nível do fator B.

NTIA – Versão 4.2.2

Programa 3: Programa NTIA para modelos com Três Fatores Hierarquizados

```
NTIA>GENESE NESTED
NTIA>NUM A B C Y;
NTIA>ARQUIVO M=ABREF(B:NEST3.DAD) A B
C Y;
NTIA>{LEIAF(M)};
NTIA>MODLIN NESTED
MOD Y = A [B(A)] B(A) [C(A B)] C(A B);
```

O *software* NTIA fornece somas de quadrados dos tipos seqüenciais e parciais, equivalentes às somas de quadrados dos tipos I e III fornecidas pelo SAS-GLM (TABELA 2). Testam, portanto, hipóteses equivalentes, exceto para a soma de quadrados do tipo parcial referente ao fator A que forneceu SQ(A) = 0,0247082 e, nesse caso, não testa a hipótese sobre as médias não ponderadas, $H_0^{(2)}$ .A soma de quadrados parciais do fator A fornecida pelo NTIA é equivalente à soma de quadrados, R [α^* | μ^*, β^*, $\gamma(\alpha\ \beta)^*$] = R [μ^*, α^*, β^*, $\gamma(\alpha\ \beta)^*$] - R [μ^*, β^*, $\gamma(\alpha\ \beta)^*$], obtida através do modelo reparametrizado proposto por Overall & Spiegel (1969) entre outros. Segundo Searle (1987), quando todas as caselas estão ocupadas (n_{ijk}>0)a soma de quadrados R [α^* | μ^*, β^*, $\gamma(\alpha\ \beta)^*$] é equivalente àquela obtida através do método dos quadrados de médias ponderadas de Yates (1934) e, portanto, testa a hipótese do tipo III sobre as médias não ponderadas. Entretanto, quando os dados são desbalanceados em presença de caselas vazias, esse procedimento falha em fornecer somas de quadrados apropriados para os testes de hipóteses do tipo III.

STATGRAPHICS – versão 7.0

O STATGRAPHICS não realiza as análises em modelos hierarquizados com dados desbalanceados.

SAEG – Versão 5.0

O SAEG fornece somas de quadrados seqüenciais equivalentes às somas de quadrados do tipo I fornecidas pelo SAS-GLM. Porém, ao contrário do SAS-GLM, considera os efeitos como aleatórios e calcula os componentes de variância.

GLIM – Versão 4.0

Programa 4: Programa GLIM para modelos com Três Fatores Hierarquizados

```
$UNITS 34 $DATA A B C Y $READ $!
1 1 1 0,791
1 1 1 0,749
    ®
2 2 2 0,915
$FACTOR A 2 B 3 C 3 $YVAR Y $FIT: +A: +B/A:
+C/B/A $!
$FINISH $!
```

O GLIM também fornece somas de quadrados do tipo seqüencial equivalentes às somas de quadrados do tipo I, fornecidas pelo SAS-GLM, sem realizar nenhum teste estatístico.

STATISTICA – Versão 5.0

O *software* STATISTICA emite a mensagem “DESIGN INCOMPLETE ; TEST PLANNED COMPARISONS OR SPECIFIC EFFECTS” e não realiza as análises. Porém, através do comando “CONTRASTS FOR BETWEEN-GROUP FACTORS”, é possível obter as somas de quadrados dos efeitos principais A, B(A) e C(A B), equivalentes às somas de quadrados do tipo IV fornecidas pelo SAS-GLM e, portanto, testam as hipóteses $H_0^{(2)}$, $H_0^{(5)}$, $H_0^{(6)}$, respectivamente.

BMDP – Versão PC90

Programa 5: Programa BMDP para modelos com Três Fatores Hierarquizados

```
/ INPUT      TITLE IS ‘MODELO
HIERARQUIZADO’.
        VARIABLES = 4.
        FORMAT = FREE.
/ VARIABLE   NAMES = A, B, C, Y.
/ BETWEEN    FACTORS = A, B, C.
        CODES(A) = 1,2.
        CODES(B) = 1 TO 5.
        CO DES(C) = 1 TO 11.
/ WEIGHTS    BETWEEN = EQUAL.
/ END
1 1 1 0,791
1 1 1 0,791
```

```
®
2 5 11 0,915
/END
ANALYSIS     PROCEDURE = STRUCTURE.
        BFORMULA = ‘A/B/C’./
END /
/ WEIGHT     BETWEEN = SIZES.
/ END
ANALYSIS     PROCEDURE = STRUCTURE.
        BFORMULA = ‘A/B/C’./

END /
```

O BMDP possui dois comandos BETWEEN = EQUAL e BETWEEN = SIZES. Se os dados são desbalanceados com todas as caselas ocupadas, então, o comando BETWEEN = EQUAL fornece somas de quadrados equivalentes às somas de quadrados do tipo III fornecidas pelo SAS-GLM. Se existem caselas vazias, as somas de quadrados fornecidas pelo BMDP são equivalentes às somas de quadrados do tipo IV do SAS-GLM. Já o comando BETWEEN = SIZES fornece somas de quadrados equivalentes às somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Testa, portanto, as hipóteses $H_0^{(1)}$, $H_0^{(4)}$, $H_0^{(6)}$, respectivamente.

CONCLUSÕES

Como foi verificado, a ocorrência de dados desbalanceados em presença de caselas vazias pode trazer sérios transtornos aos pesquisadores das ciências aplicadas, em relação às interpretações de hipóteses estatísticas pois, na maioria dos casos, a falta de uma documentação explícita sobre o que esses *softwares* estão calculando, pode induzir a tomada de decisões incorretas, comprometendo o "resultado" de suas pesquisas.

Sendo assim, os pesquisadores, usuários de *softwares* estatísticos, devem ser cautelosos na análise estatística de dados desbalanceados, evitando o uso indiscriminado de *softwares* estatísticos sem o conhecimento prévio de sua documentação. Considera-se, portanto, de vital importância o acompanhamento de um profissional da estatística, tanto no planejamento do experimento, quanto na análise dos dados e na interpretação dos resultados.

Face aos resultados obtidos, concluiu-se que:

• As somas de quadrados do tipo I fornecidas pelo SAS-GLM, para os fatores A, B(A) e C(A B), correspondem às hipóteses do tipo I sobre as médias ponderadas não ajustadas e testam as hipóteses $H_0^{(1)}$, $H_0^{(4)}$, $H_0^{(6)}$, respectivamente.

• As somas de quadrados do tipo II fornecidas pelo SAS-GLM, não testam as hipóteses do tipo II sobre as médias ponderadas ajustadas, como nos casos de esquemas fatoriais, pois o fator C está aninhado sob o fator B, onde o fator B está aninhado sob o fator A. Nesse caso, as hipóteses associadas às somas de quadrados do tipo II são equivalentes às hipóteses do tipo I.

• Quando os níveis do fator C(A B) são diferentes, independentemente dos dados serem balanceados ou não, a soma de quadrados do tipo III referente ao fator A, fornecida pelo SAS-GLM não testa a hipótese, $H_0^{(2)}$, mas uma hipótese gerada a partir de funções estimáveis complexas do tipo III, $H_0^{(3)}$.

• Dos *softwares* estudados, apenas o STATGRAPHICS não realiza as análises para modelos com dados desbalanceados.

• O MINITAB, o SAEG, o GLIM e o NTIA fornecem em suas saídas somas de quadrados do tipo seqüencial, equivalentes às somas de quadrados do tipo I do SAS-GLM. Deve-se, entretanto, ressaltar que o MINITAB e o GLIM não realizam nenhum teste estatístico e o SAEG, considera os efeitos aleatórios e calcula os componentes de variância.

• O NTIA fornece também as somas de quadrados do tipo parcial, mas apenas as somas de quadrados referentes aos fatores B(A) e C(A B) são equivalentes às somas de quadrados do tipo III fornecidas pelo SAS-GLM. A soma de quadrados referente ao fator A não testa a hipótese sobre as médias não ponderadas, $H_0^{(2)}$, e nem a hipótese $H_0^{(3)}$ testada pelo SAS-GLM.

• O STATISTICA fornece somas de quadrados para os efeitos principais, A, B(C), e C(A B), equivalentes às somas de quadrados do tipo IV fornecidas pelo SAS-GLM. Nesse caso, testam as hipóteses $H_0^{(2)}$, $H_0^{(4)}$, $H_0^{(6)}$.

• Se os dados são desbalanceados com todas as caselas ocupadas, o comando BETWEEN = EQUAL do BMDP fornece somas de quadrados equivalentes às somas de quadrados do tipo III do SAS-GLM. Agora, se existem caselas vazias, então, elas fornecem somas de quadrados equivalentes às somas de quadrados do tipo IV do SAS-GLM.

• Já o comando BETWEEN = SIZES do BMDP fornece somas de quadrados equivalentes às somas de quadrados dos tipos I e II do SAS-GLM e, portanto, testam hipóteses do tipo I sobre as médias ponderadas não ajustadas.

REFERÊNCIAS BIBLIOGRÁFICAS

DALLAL, G.E. The computer analysis of factorial experiments with nested factors. **The American Statistician**, v.46, p.240, 1992.

ELLIOTT, A.C.; WOODWARD, W. A. Analysis of an unbalanced two-way anova on the microcomputer. **Communications in Statistics - Simulations**, v.15, p.215-225, 1986.

FRANCIS, I.A comparison of several analysis of variance programs. **Journal of the American Statistical Association**, v.68, p.860-865, 1973.

HERR, D.G. On the history of anova in unbalanced, fatorial designs: The first 30 years. **The American Statistician**, v.40, p.265-270, 1986.

HOCKING, R.R. **The analysis of linear models**. Monterey, California: Brooks/Cole Publishing Company, 1985. 385p.

IEMMA, A.F. **Modelos lineares:** uma introdução para profissionais da pesquisa agropecuária. Londrina: Imprensa Oficial do Estado do Paraná, 1987. 263p.

IEMMA, A.F. **Análisis de varianza con datos desbalanceados**. Bogotá: Universidad Nacional de Colombia, 1993. 120p.

IEMMA, A.F. Que hipóteses estatísticas testamos através do "SAS" em presença de caselas vazias? ***Scientia Agricola***, v.52, p.210-220,1995/a.

IEMMA, A.F. Análise de variância de dados desbalanceados. In: CONGRESSO BRASILEIRO DE USUÁRIOS DO SAS, 4., Piracicaba, 1995. **Anais**. ESALQ/USP, 1995b. 111p.

IEMMA, A.F. Dados estatísticos desbalanceados: PROC SAS/GLM. In: SEMANA DE MATEMÁTICA, 2., Rio de Janeiro, 1995. **Anais**. UFRJ, 1995c. 130p.

IEMMA, A.F. **Analisis de varianza de experimentos con celdas vazias**. Cordoba: Universidad Nacional de Cordoba, 1997. 112p. (Trabajos de Matematicas – Serie C – N. 22/97).

IEMMA, A.F.; PERRI, S. H. V. **Ajuste de modelos mistos desbalanceados através do sistema estatístico SAS**. Piracicaba: Departamento de Matemática e Estatística, ESALQ, USP. 1997. 99p.

MONDARDO, M.; IEMMA, A.F. Sobre quatro tipos de funções estimáveis fornecidas pelo PROC GLM do SAS para dados desbalanceados. ***Scientia Agricola***, v.55, p.172-182, 1998.

OVERALL, J.E.; SPIEGEL, D.K. Concerning least squares analysis of experimental data. **Psichological Bulletin**, v.72, p.311-322, 1969.

PADOVANI, C.R. Estimabilidade no modelo linear em classificação hierárquica com s estágios. Piracicaba, 1984, 81p. Tese (Doutorado) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

RAO, C.R. On the linear combination of observations and the general theory of least squares. **Sankhyã**, v.7, p.237-256, 1945.

SANTOS, E.S. Utilização de "*Softwares*" estatísticos na interpretação de hipóteses com dados desbalanceados. Piracicaba, 1994, 175p. Tese (Doutorado) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

SEARLE, S.R. **Linear models for unbalanced data**. New York: John Wiley, 1987. 536p.

SEARLE, S.R. Analysis of variance computing package output for unbalanced data from fixed-effects models with nested factors. **The American Statistician**, v.48, p.148-153, 1994.

WINER, B.J. **Statistical principles in experimental design**. 2.ed. New York: McGraw- Hill Book, 1971. 907p.

YATES, F. The principles of orthogonality and confounding in replicated experiments. **Journal of Agricultural Science**, v.23, p.108-145, 1933.

YATES, F. The analysis of multiple classifications with unequal numbers in the different classes. **Journal of the American Statistical Association**, v.29, p.51-66, 1934.

Recebido para publicação em 31.07.98
Aceito para publicação em 26.07.99