EMPREGO DO MODELO SUPERPARAMETRIZADO EM EXPERIMENTO FATORIAL DESBALANCEADO COM DOIS FATORES[1]

Overparameterized model for an unbalanced factorial experiment with two factors

Eliana Mara Manso[2], Augusto Ramalho de Morais[3]

RESUMO

Na pesquisa agropecuária é comum o estudo de vários fatores e freqüentemente ocorrem perdas de observações, constituindo assim um experimento desbalanceado. É necessário conhecer as hipóteses testadas através dos sistemas estatísticos e ocorrendo caselas vazias a interpretação é ainda mais complexa, pois geralmente, as hipóteses sobre os efeitos principais de um dos fatores contêm os efeitos principais de outros fatores e os efeitos de interações. Adotando o modelo superparametrizado, com este trabalho, objetivou-se desenvolver esquemas de análises de variâncias de dados desbalanceados e/ou com caselas vazias, identificar e interpretar as hipóteses associadas às somas de quadrados através do procedimento General Linear Models (GLM) do Statistical Analysis System (SAS), que provêm quatro tipos de somas de quadrados. Foram analisados dois casos distintos, utilizando dados referentes ao peso comercial de cenoura, provenientes de experimento inteiramente ao acaso, tendo como fatores cultivares e fases da lua como épocas de plantio. Em face aos resultados obtidos, verificou-se que, quando os dados são desbalanceados, as funções estimáveis de um fator envolvem os parâmetros relativos ao fator e os componentes das interações nas quais o fator está presente; as somas de quadrados do tipo III equivalentes as do tipo IV e a ordenação dos fatores principais não afeta as hipóteses do tipo I. Entretanto, quando ocorreram caselas vazias no modelo com dois fatores, os quatro tipos de somas de quadrados para o fator principal de entrada foram diferentes e; a ordenação é fundamental para obtenção das hipóteses do tipo I. Quando ocorrem perdas de parcelas, a identificação das funções estimáveis é complexa e as hipóteses ficam de difícil interpretação. Nas funções estimáveis de interações ocorrem parâmetros da própria interação. Diferenças entre níveis do fator A somente podem ser estimados na presença de efeitos médios do fator B e da interação.

Termos para indexação: Modelo linear, somas de quadrados, dados desbalanceados, caselas vazias.

ABSTRACT

In agricultural research it is common to study simultaneously various factors and natural loss of observations frequently leads to unbalanced experiments. Thus it is necessary to know which hypothesis can be tested in the statistical systems. In a missing cells scenario the interpretation is even more complex. In general, the hypothesis on main effects of one of these factors contains the main effects of other factors and of effects of interactions. The aim of this work is to develop ANOVA schemes for the overparameterized model to unbalanced data and, or with situations with missing cells. Additionally we call attention to correct identification and interpretation of the hypothesis associated with the four types of sum of squares given by SAS-GLM procedure. Two distinct cases were analyzed, namely: using data referring to commercial weight of carrots, arising from a completely randomized experiment and, using a factorial design with two planting factors (cultivation and the phasis of the moon). We conclude that the estimable functions of a factor involves a linear combination of both main effects and interaction parameters for both Type III and Type IV Sum of Squares. The order of the main effects changes type I Sum of Squares. Thus, when there are missing cells in the two-factor model, the four types of sum of square for main effects are different and the order is fundamental to obtain the correct type I hypothesis. When missing cells happens, the identification of the estimable functions is more complex and the hypotheses are difficult to interpret. In the estimable functions for interaction, interaction parameters appears as expected. In this case, differences between levels of one main effect factor can only be estimated in the presence of average effects due to the other factor B and the interaction.

Index terms: Linear model, Missing Cell, Sum of Squares, Unbalanced data.

(Recebido para publicação em 27 de abril de 2004 e aprovado em 18 de outubro de 2005)

INTRODUÇÃO

Na estatística experimental, principalmente na pesquisa em agropecuária, a análise de variância de fatoriais com número de repetições constantes, caracterizando um delineamento balanceado, é relativamente fácil e amplamente conhecida, porém quando ocorre um número diferente de repetições e/ou caselas vazias, caracterizando um delineamento desbalanceado, torna-se mais complexa. Um outro problema que ocorre na presença de caselas vazias é que, ao adotar-se um modelo superparametrizado, o número de parâmetros pode ser maior do que o número de caselas disponíveis para estimá-los, influenciando a formulação de hipóteses, que podem envolver parâmetros

[1] Projeto financiado pelo CNPq.

[2] Mestranda em Agronomia, Estatística e Experimentação Agropecuária – Universidade Federal de Lavras/UFLA – Cx. P. 3037 – 37200-000 – Lavras, MG – Bolsista do CNPq – elianamanso@bol.com.br

[3] Professor Adjunto, Departamento de Ciências Exatas da Universidade Federal de Lavras/UFLA – Cx. P. 3037 – 37200-000 – Lavras, MG – armorais@ufla.br

sem interesse, induzindo a hipóteses equivocadas. Nos últimos vinte e cinco anos, o surgimento de aplicativos estatísticos facilitou a análise de fatoriais desbalanceados, porém ocorrem diferenças nas saídas das somas de quadrados entre os sistemas computacionais, acarretando dúvidas nas interpretações.

A base para início da revolução computacional nos anos sessenta e para os métodos de análise usados atualmente é encontrada nos artigos originais de Yates (1933, 1934). Yates (1934) estabelece três métodos para cálculo das somas de quadrados: método das médias não ponderadas; método das médias dos quadrados ponderados; método do ajuste de constantes, iniciando a parametrização sucessiva, de grande utilidade quando as somas de quadrados são estudadas através da notação R( ).

Um modelo linear é dito superparametrizado quando explicita um parâmetro para cada efeito dos fatores envolvidos (SEARLE, 1971). Partindo de um modelo mais simples até o modelo com dois fatores fixos e com interação, tem-se:

$$y = X_1\theta_1 + e_1 \quad \Rightarrow \quad y_i = \mu + e_i \qquad (S.1)$$
$$y = X_2\theta_2 + e_2 \quad \Rightarrow \quad y_{ik} = \mu + \alpha_i + e_{ik} \qquad (S.2)$$
$$y = X_3\theta_3 + e_3 \quad \Rightarrow \quad y_{jk} = \mu + \beta_j + e_{jk} \qquad (S.3)$$
$$y = X_4\theta_4 + e_4 \quad \Rightarrow \quad y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \qquad (S.4)$$
$$y = X_5\theta_5 + e_5 \quad \Rightarrow \quad y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad (S.5)$$

em que X_1, X_2, X_3, X_4 e X_5 são matrizes do planejamento de cada modelo (S), y é vetor das observações, e_1, e_2, e_3, e_4 e e_5 são os vetores dos erros, e θ_1, θ_2, θ_3, θ_4 e θ_5 são os vetores de parâmetros dos modelos S.1, S.2, …, S.5, respectivamente.

Esses modelos permitem a identificação de todos os efeitos; no entanto, quando ocorre dados desbalanceados alguns problemas podem ocorrer. As discrepâncias entre somas de quadrados resultantes de diferenças entre as hipóteses testadas foram alertadas por Searle (1987) e por Speed et al. (1978), que fazem uma descrição detalhada dos métodos de análises e relatam que o SAS incorporou parte desses métodos ao procedimento General Linear Models (GLM).

Speed & Hocking (1976) relatam as idéias de parametrização sucessiva por meio do modelo superparametrizado. As parametrizações sucessivas e ordenadas facilitam a interpretação da notação R(.) e de certas somas de quadrados a elas associadas. O termo R(.) pode ser interpretado como medida de variação em y explicada pelo modelo ajustado.

Searle et al. (1980) discorrem sobre os quatro tipos de funções estimáveis fornecidas pelo PROC GLM do SAS. As somas de quadrados são interpretadas por meio das funções estimáveis, sendo mais informativas que os valores fornecidos por R( ): Tipo I: valores R( ) são seqüenciais; Tipo II: valores R( ) para cada fator ajustado a todos os outros que não o contém; Tipo III: valores R( ) para cada fator ajustado a todos os outros; Tipo IV: fixa uma hipótese e testa-a pela estatística F; contrário aos tipos anteriores, não pretende proporcionar funções estimáveis para explicarem a soma de quadrados pré-definida.

A perda de observações e de caselas inteiras, com suas implicações nas funções estimáveis e na formulação de hipóteses foi estudada por Freund (1980), que compara os procedimentos e a análise dos dados por dois métodos: pelo PROC GLM do SAS que usa a metodologia das funções estimáveis; e, pelo método da reparametrização.

Herr (1986) faz uma descrição dos primeiros trinta anos da Anova em fatorial com dados desbalanceados.

Searle (1987) estabelece relação entre os tipos de somas dos quadrados do SAS, caracterizadas como seqüencial, ajuste para todos os fatores e interações, exceto aquelas que envolvem o fator de interesse, ajuste para todos os efeitos envolvidos no modelo com restrição paramétrica do tipo å e ajuste envolvendo hipóteses ou parte das caselas. Ressalta que as somas de quadrados caracterizam-se como:

- Tipo I = II = III = IV quando os dados são balanceados;
- Tipo II = III = IV para modelos sem interação;
- Tipo III = IV para os modelos com todas as caselas completas.

O modelo superparametrizado (modelo-S) é excelente para interpretação prática e para utilização da notação R( ). De acordo com Searle (1987), a redução da soma de quadrados para ajuste do modelo (S.5), é dada por:

$R(\theta) = y'X(X'X)^{-}X'y = \theta^{0}'X'y$ = S.Q. Parâmetros (1)
sendo θ^{0} uma solução para $X'X\theta = X'y$.

A estimabilidade de funções paramétricas com dados desbalanceados foi avaliada por Mondardo (1994), a fim de auxiliar a escolha da melhor opção dentre as somas de quadrados da saída do procedimento GLM do SAS. Relata que modelo superparametrizado, todos os tipos de hipóteses envolvem os parâmetros referentes à interação, não se podendo testar os efeitos principais isoladamente, somente a hipótese sobre a interação é livre de parâmetros sem interesse. A soma de quadrados Tipo I para efeitos principais, não apresenta aparente interesse porque pode envolver parâmetros do outro fator ou mesmo em conjuntos completos não ortogonais de modelos sem interação, além de parâmetros da interação nos modelos com interação. A Tipo II, apesar de testar hipóteses de efeitos principais ajustados, sempre contém combinações lineares dos parâmetros relativos à interação e coeficientes pouco usuais. Na do Tipo III, se aumentar o número de caselas vazias e o número de níveis do fator também, ela se torna complexa e foge ao interesse do pesquisador. A soma de quadrados Tipo IV testa hipóteses bem mais simples e fáceis de interpretar, entretanto com a perda de informação, só é considerada uma parte dos dados com os quais elas podem trabalhar como se fossem dados balanceados.

Maiores detalhes dos testes de hipóteses e aplicativos podem ser vistos em Camarinha Filho (1995) e Santos (1994).

Com base num modelo com dois fatores de efeitos fixos, em presença ou não de caselas vazias, Iemma (1995) apresenta as hipóteses mais comuns sobre os efeitos de linhas, colunas e interação e comenta que as somas de quadrados do Tipo I, fornecidas pelo PROC GLM do SAS, para dados desbalanceados, dados que são obtidas seqüencialmente, dependem da ordem de entrada dos parâmetros no modelo.

Wechsler (1998) enfatiza os perigos de empregar os aplicativos estatísticos sem antes conhecer as hipóteses por eles testadas em fatoriais fixos desbalanceados. Considera que o modelo superparametrizado não define claramente o que se entende por efeito dos fatores principais e efeito da interação, pois os parâmetros relativos a esses efeitos não são individualmente estimáveis, contanto que alguma restrição seja adotada (SEARLE, 1971), como as restrições usuais: $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_i = \sum_{i=1}^{a}\gamma_{ij} = \sum_{j=1}^{b}\gamma_{ij} = 0$

Relata ainda que, num fatorial desbalanceado as somas de quadrados para os efeitos principais e interação, bem como as hipóteses a elas associadas podem variar, dependendo das restrições e procedimentos computacionais empregados, causando mal-entendidos sobre as hipóteses testadas pelos aplicativos.

Visa-se com este artigo, apresentar esquemas de análises de variâncias para experimentos com dois fatores, com dados desbalanceados e/ou caselas vazias; desenvolver procedimentos para análise desses dados, utilizando o sistema computacional SAS e identificar e interpretar as hipóteses associadas às somas de quadrados, com finalidade de proporcionar um alerta aos usuários.

MATERIAL E MÉTODOS

Para ilustrar os procedimentos apresentados, utiliza-se dados sobre peso comercial de raízes de cenoura, provenientes de experimento realizado no setor de Olericultura do Departamento de Agricultura da Universidade Federal de Lavras, UFLA, em Lavras, Minas Gerais.

Nesse experimento foram avaliados as cultivares: Kuronan, Carandaí e Brasília (nacionais) e Nantes (importado) e as épocas de plantio: fases da lua (Crescente, Cheia, Minguante e Nova) no mês de julho. A parcela experimental era de 2 m de comprimento por 1 m de largura. Os tratos culturais foram realizados de acordo com a necessidade da cultura.

O modelo linear superparametrizado, assume a caracterização:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \quad (2)$$

em que, i = 1, 2, 3 e 4; j = 1, 2, 3 e 4; k = 0, ..., n_{ij} repetições;
y_{ijk} : observação referente a i-ésima cultivar na j-ésima lua e na k-ésima repetição;
μ : constante em todas as observações;
α_i: efeito da i-ésima cultivar (fator A);
β_j: efeito da j-ésima fase da lua (fator B);
γ_{ij}: efeito da interação entre o i-ésimo nível do fator A e o j-ésimo nível do fator B;
e_{ijk} : erro experimental associado à observação y_{ijk}, considerado independente e normalmente distribuído com média zero e variância constante, tais que $e_{ijk} \sim N(0, s^2)$.

- Para a ordenação A-B: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$.
- Para a ordenação B-A: $y_{jik} = \mu + \beta_j + \alpha_i + \gamma_{ij} + e_{jik.}$

1º caso: fatorial 4 x 4 desbalanceado em relação ao número de repetições

Na Tabela 1, estão os resultados obtidos da produção comercial das raízes de diferentes cultivares de cenoura nas diversas fases da lua.

TABELA 1 – Produção comercial de raízes de cultivares de cenoura, em t/ha, de acordo com as diferentes fases da lua (1º caso).

Fator A (cultivares)	Fator B (fases da lua)			
	nova	crescente	cheia	minguante
Nantes	37,55	33,85	37,75	27,15
	41,45	x	28,85	19,80
	18,15	x	30,40	15,20
Kuronan	50,00	51,80	42,90	31,70
	44,85	33,55	x	38,95
	52,45	21,40	35,65	23,65
Carandaí	49,70	48,75	35,60	39,85
	45,90	17,95	29,95	28,05
	31,00	22,95	38,55	31,80
Brasília	56,00	55,05	55,25	x
	54,85	30,70	38,75	x
	54,15	29,15	44,25	50,65

Fonte: Cristiane R. B. Aguirre Ramos. Departamento de Agricultura. UFLA, Lavras/MG.
Nota: x: dado omitido propositadamente para este estudo.

2º caso: fatorial 4 x 4 desbalanceado em relação às combinações dos fatores

Na Tabela 2, reproduz-se dados do experimento do primeiro caso, porém com desbalanceamento e com caselas vazias.

TABELA 2 – Produção comercial de raízes de cultivares de cenoura, em t/ha, de acordo com as diferentes fases da lua (2º caso).

Fator A (cultivares)	Fator B (fases da lua)			
	Nova	crescente	cheia	minguante
Nantes	37,55	x	37,75	27,15
	41,45	x	28,85	19,80
	18,15	x	30,40	15,20
Kuronan	50,00	51,80	42,90	31,70
	44,85	33,55	x	38,95
	52,45	21,40	35,65	23,65
Carandaí	49,70	48,75	35,60	39,85
	45,90	17,95	29,95	28,05
	31,00	22,95	38,55	31,80
Brasília	56,00	55,05	55,25	x
	54,85	30,70	38,75	x
	54,15	29,15	44,25	x

Fonte: Cristiane R. B. Aguirre Ramos. Departamento de Agricultura. UFLA, Lavras/MG.
Nota: x: dado omitido propositadamente para este estudo.

Procedimento GLM do Sistema Estatístico SAS

Para realização das análises de variância utilizou-se o procedimento GLM do SAS, conforme SAS Institute (1990). O programa SAS GLM para modelos com dois fatores em esquema fatorial é:

```
data caso1;
input  A  B  peso;           /* leitura das informações */
cards;                       /* início de linhas de dados na ordem dada:
                                nível de A, nível de B e variável resposta */
1      1      37.55
1      1      41.45
1      1      18.15
1      2      33.85
1      2      .              /* indica valor ausente */
1      2      .
1      3      37.75
1      3      28.85
1      3      30.40
.
.
4      3      38.75
4      3      44.25
4      4      .
4      4      .
4      4      50.65
;
proc print data=caso1;       /* gera e imprime um relatório */
run;
proc glm;                    /* análise de variância para dados desbalanceados*/
  class A B;                 /* ordem de entrada dos dados: A, B, A-B */
  model peso=A B A*B/E E1 E2 E3 E4 SS1 SS2 SS3 SS4 XPX I;
run;
proc glm;
run;                         /* fim de seção de programa */
quit;                        /* encerra o programa */
```

RESULTADOS E DISCUSSÃO

1º caso: fatorial 4 x 4 desbalanceado em relação ao número de repetições

As funções estimáveis que estão associadas às hipóteses testadas, no Modelo-S, são obtidas atribuindo-se valor um para cada um dos coeficientes e zerando os demais, conforme sugestão de Mondardo (1994). Na presença de interação não obtiveram-se funções estimáveis exclusivamente sobre os efeitos principais, ocorrendo normalmente, um contraste entre os níveis de um fator, seguido de outros parâmetros.

As funções estimáveis para o fator A que vão constituírem as hipóteses testadas pelas quatro somas de quadrados fornecidas pelo Proc GLM do SAS, são:

Tipo I $\Rightarrow$ H_0^1

$\alpha_1 - \alpha_4 + 1/30\ (-6\beta_2 + 6\beta_4 + 10\gamma_{11} + 3\gamma_{12} + 10\gamma_{13} + 10\gamma_{14} - 10\gamma_{41} - 10\gamma_{42} - 10\gamma_{43} - 3\gamma_{44})$
$\alpha_2 - \alpha_4 - 0{,}0273\ (\beta_1+\beta_2) - 0{,}1182\beta_3 + 0{,}1727\beta_4 + 1/11(3\gamma_{21} + 3\gamma_{22} + 2\gamma_{23} + 3\gamma_{24}) - 1/3(\gamma_{41} + \gamma_{42} - \gamma_{43}) - 1/5\gamma_{44}$
$\alpha_3 - \alpha_4 - 1/20\ (\beta_1 - \beta_2 - \beta_3 - 3\beta_4 + 1/12(3\gamma_{31} + 3\gamma_{32} + 3\gamma_{33} + 3\gamma_{34} - 4\gamma_{41} - 4\gamma_{42} - 4\gamma_{43}) - 1/5\gamma_{44}$

Tipo II $\Rightarrow$ H_0^2

$\alpha_1 - \alpha_4 + 1/16\ (5\gamma_{11} + 2\gamma_{12} + 5\gamma_{13} + 4\gamma_{14} + \gamma_{21} - \gamma_{24} + \gamma_{32} - \gamma_{34} - 5\gamma_{41} - 4\gamma_{42} - 5\gamma_{43} - 2\gamma_{44})$
$\alpha_2 - \alpha_4 + 0{,}0102\gamma_{11} + 0{,}0041\gamma_{12} + 0{,}036\gamma_{13} + 0{,}0503\gamma_{14} + 0{,}284\gamma_{21} + 0{,}286\gamma_{22} + 0{,}2065\gamma_{23} + 0{,}2235\gamma_{24} + 0{,}082\gamma_{31} + 0{,}0102\gamma_{32} + 0{,}034_{33} - 0{,}0523\gamma_{34} - 0{,}3023\gamma_{41} - 0{,}3003\gamma_{42} - 0{,}2765\gamma_{43} - 0{,}1209\gamma_{44}$
$\alpha_3 - \alpha_4 + 0{,}017\gamma_{11} + 0{,}0068\gamma_{12} + 0{,}0183\gamma_{13} - 0{,}0421\gamma_{14} + 0{,}0149\gamma_{21} + 0{,}0183\gamma_{22} + 0{,}0109\gamma_{23} - 0{,}0442\gamma_{24} + 0{,}2636\gamma_{31} + 0{,}267\gamma_{32} + 0{,}2649_{33} - 0{,}2045\gamma_{34} - 0{,}2955\gamma_{41} - 0{,}2921\gamma_{42} - 0{,}2942\gamma_{43} - 0{,}1182\gamma_{44}$

Tipo III = Tipo IV $\Rightarrow$ $H_0^3 = H_0^4$

$\alpha_1 - \alpha_4 + 1/4\ (\gamma_{11} + \gamma_{12} + \gamma_{13} + \gamma_{14} - \gamma_{41} - \gamma_{42} - \gamma_{43} - \gamma_{44})$
$\alpha_2 - \alpha_4 + 1/4\ (\gamma_{21} + \gamma_{22} + \gamma_{23} + \gamma_{24} - \gamma_{41} - \gamma_{42} - \gamma_{43} - \gamma_{44})$
$\alpha_3 - \alpha_4 + 1/4\ (\gamma_{31} + \gamma_{32} + \gamma_{33} + \gamma_{34} - \gamma_{41} - \gamma_{42} - \gamma_{43} - \gamma_{44})$

Por meio das funções estimáveis pode-se identificar a hipótese que está sendo testada e, nota-se certa dificuldade nessa interpretação. Para o fator A, cada hipótese do Tipo I, apresenta um contraste entre níveis do efeito principal, seguida de um contraste entre níveis do outro fator, devido ao fato que nenhum ajuste é feito e acompanhada de efeitos de parâmetros de interações. As hipóteses do Tipo II não trazem contraste do outro fator, apenas da interação, mas seus coeficientes não são de natureza prática. As hipóteses do Tipo III e IV são equivalentes e de aparente interesse por apresentarem contraste entre os níveis do fator A e coeficientes mais simples para a interação, podendo ser a hipótese mais adequada quando pretende-se testar efeitos de cada fator envolvido, confirmando resultado de Santos (1994).

Para este conjunto de dados, sem caselas vazias, a hipótese sobre a interação tem (a–1)(b–1) = (4–1)(4–1) = 9 números de graus de liberdade e, conseqüentemente, nove funções paramétricas linearmente independentes de efeitos de interação poderão ser testadas. Essas hipóteses são livres de parâmetros sem interesse dos fatores principais.

Como os dados são desbalanceados, a ordem de entrada dos fatores no modelo foi considerada na formulação das hipóteses e somas de quadrados associadas. Assim, com os resultados das análises de variância (Tabela 3), tem-se que a ordem de entrada dos efeitos dos fatores principais no modelo, afeta as somas de quadrados tipo I, que são obtidas seqüencialmente, isto é, $R(\alpha|\mu) \neq R(\alpha|\mu,\beta)$ e $R(\beta|\mu) \neq R(\beta|\mu,\alpha)$. Para os demais tipos de somas de quadrados, a ordem dos fatores no modelo não altera o resultado.

As somas de quadrados do Tipo II, eqüivalem as somas de quadrados ajustadas para todos os fatores e interações, exceto interações e/ou fatores hierarquizados que envolvem o fator de interesse, ou seja, $R(\alpha|\mu,\beta)$ e $R(\beta|\mu,\alpha)$. As funções paramétricas estimáveis incluem contrastes de difícil interpretação prática e a ordem de entrada dos fatores no modelo não afetou os resultados das somas de quadrados.

Com referência às somas de quadrados associadas às hipóteses dos tipos III e IV constata-se que são equivalentes, confirmando afirmativa de Iemma (1995), Searle (1987) e Speed et al. (1978) que este fato ocorre quando se trata de dados desbalanceados e sem caselas vazias.

TABELA 3 – Análise de variância para os dados da Tabela 1, fornecida pelo PROC GLM do SAS,com base no modelo $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}$

Ordem A-B					Ordem B-A				
C.V	G.L	H_0	R( )	SS	C.V	G.L	H_0	R( )	SS
TIPO I									
A	3	H_0^1	R(α|μ)	1692,384	B	3	H_0^5	R(β|μ)	1172,914
B(aj)	3	H_0^6	R(β|μ,α)	1091,033	A(aj)	3	H_0^2	R(α|μ,β)	1610,503
A-B	9	H_0^9	R(γ|μ,α,β)	397,030	B-A	9	H_0^9	R(γ|μ,α,β)	397,030
TIPO II									
A(aj)	3	H_0^2	R(α|μ,β)	1610,503	B(aj)	3	H_0^6	R(β|μ,α)	1091,033
B(aj)	3	H_0^6	R(β|μ,α)	1091,033	A(aj)	3	H_0^2	R(α|μ,β)	1610,503
A-B	9	H_0^9	R(γ|μ,α,β)	397,030	B-A	9	H_0^9	R(γ|μ,α,β)	397,030
TIPO III									
A	3	H_0^3	R(α̇|μ̇,β̇,γ̇)	1366,979	B	3	H_0^7	R(β̇|μ̇,α̇,γ̇)	754,253
B	3	H_0^7	R(β̇|μ̇,α̇,γ̇)	754,253	A	3	H_0^3	R(α̇|μ̇,β̇,γ̇)	1366,979
A-B	9	H_0^9	R(γ̇|μ̇,α̇,β̇)	397,030	B-A	9	H_0^9	R(γ̇|μ̇,α̇,β̇)	397,030
TIPO IV									
A	3	H_0^4	—	1366,979	B	3	H_0^4	—	754,253
B	3	H_0^8	—	754,253	A	3	H_0^8	—	1366,979
A-B	9	H_0^9	—	397,030	B-A	9	H_0^9	—	397,030

2º caso: fatorial 4 x 4 desbalanceado em relação às combinações dos fatores

Quando existe a presença de caselas vazias, para o Modelo-S com interação, constata-se que não há possibilidade de obterem-se funções paramétricas estimáveis exclusivamente sobre os efeitos principais.

As funções paramétricas estimáveis para o fator A fornecidas pelo Proc GLM do SAS, que constituem um conjunto de hipóteses testadas pelas quatro somas de quadrados são:

Tipo I $\Rightarrow H_0^1$

$\alpha_1 - \alpha_4 + 1/3\ (-\beta_2 + \beta_4 + \gamma_{11} + \gamma_{13} + \gamma_{14} - \gamma_{41} - \gamma_{42} - \gamma_{43})$

$\alpha_2 - \alpha_4 + 1/33\ (-2\beta_1 - 2\beta_2 + 5\beta_3 + 9\beta_4 + 9\gamma_{21} + 9\gamma_{22} + 6\gamma_{23} - 9\gamma_{24} - 11\gamma_{41} + 11\gamma_{42} - 11\gamma_{43})$

$\alpha_3 - \alpha_4 + 1/12\ (-\beta_1 - \beta_2 - \beta_3 + 3\beta_4 + 3\gamma_{31} + 3\gamma_{32} + 3\gamma_{33} - 3\gamma_{34} - 4\gamma_{41} + 4\gamma_{42} - 4\gamma_{43})$

Tipo II $\Rightarrow H_0^2$

$\alpha_1 - \alpha_4 + 1/8\ (3\gamma_{11} + 3\gamma_{13} + 2\gamma_{14} + \gamma_{22} - \gamma_{24} + \gamma_{32} - \gamma_{34} - 3\gamma_{41} - 2\gamma_{42} - 2\gamma_{43})$

$\alpha_2 - \alpha_4 + 0{,}0294\gamma_{11} + 0{,}0564\gamma_{13} - 0{,}0858\gamma_{14} + 0{,}2966\gamma_{21} + 0{,}3064\gamma_{22} + 0{,}2157\gamma_{23} + 0{,}1814\gamma_{24} + 0{,}0196\gamma_{31} + 0{,}0294\gamma_{32} + 0{,}0466\gamma_{33} - 0{,}0956\gamma_{34} - 0{,}3456\gamma_{41} - 0{,}3358\gamma_{42} - 0{,}3186\gamma_{43}$

$\alpha_3 - \alpha_4 + 0{,}0368\gamma_{11} + 0{,}0392\gamma_{13} - 0{,}076\gamma_{14} + 0{,}027\gamma_{21} + 0{,}0392\gamma_{22} + 0{,}0196\gamma_{23} - 0{,}0858\gamma_{24} + 0{,}2745\gamma_{31} + 0{,}2868\gamma_{32} + 0{,}277\gamma_{33} + 0{,}1618\gamma_{34} - 0{,}3382\gamma_{41} - 0{,}326\gamma_{42} - 0{,}3358\gamma_{43}$

Tipo III $\Rightarrow$ H_0^3

$\alpha_1 - \alpha_4 + 1/8\ (3\gamma_{11} + 3\gamma_{13} + 2\gamma_{14} + \gamma_{22} - \gamma_{24} + \gamma_{32} - \gamma_{34} - 3\gamma_{41} - 2\gamma_{42} - 3\gamma_{43})$

$\alpha_2 - \alpha_4 + 1/80\ (3\gamma_{11} + 3\gamma_{13} - 6\gamma_{14} + 22\gamma_{21} + 23\gamma_{22} + 22\gamma_{23} + 13\gamma_{24} + 2\gamma_{31} + 3\gamma_{32} + 2\gamma_{33} - 7\gamma_{34} - 27\gamma_{41} - 26\gamma_{42} - 27\gamma_{43})$

$\alpha_3 - \alpha_4 + 1/80\ (3\gamma_{11} + 3\gamma_{13} - 6\gamma_{14} + 2\gamma_{21} + 3\gamma_{22} + 2\gamma_{23} - 7\gamma_{24} + 22\gamma_{31} + 23\gamma_{32} + 22\gamma_{33} + 13\gamma_{34} - 27\gamma_{41} - 26\gamma_{42} - 27\gamma_{43})$

Tipo IV $\Rightarrow$ H_0^4

$\alpha_1 - \alpha_4 + 1/2\ (\gamma_{11} + \gamma_{13} - \gamma_{41} - \gamma_{43})$

$\alpha_2 - \alpha_4 + 1/3\ (\gamma_{21} + \gamma_{22} + \gamma_{23} - \gamma_{41} - \gamma_{42} - \gamma_{43})$

$\alpha_3 - \alpha_4 + 1/3\ (\gamma_{31} + \gamma_{32} + \gamma_{33} - \gamma_{41} - \gamma_{42} - \gamma_{43})$

A hipótese Tipo I para o fator A, H_0^1, envolve além dos parâmetros α, também os parâmetros β e γ, sendo a soma de quadrados Tipo I equivalente a $R(\alpha|\mu)$, não ajustada para o fator B e interação. É de difícil interpretação, tornando-se praticamente impossível visualizar o que está sendo testado. Contém coeficientes confusos, pois é uma hipótese sobre médias ponderadas de linhas não ajustadas para colunas, em diferentes freqüências.

A interpretação das hipóteses apresenta dificuldade. Todos os tipos de hipóteses envolvem os parâmetros referentes à interação, portanto não se podem testar isoladamente os efeitos principais. Com exceção da hipótese sobre a interação, as demais tem componentes do outro fator e/ou interação. Na presença de interação não se obtêm funções estimáveis exclusivamente sobre os efeitos principais ocorrendo, normalmente um contraste entre os níveis de um fator, seguido de outros parâmetros, ou seja, os efeitos do outro fator e/ou de interações, reforçando o alerta de Camarinha Filho (1995), que o problema principal não está apenas na interpretação das hipóteses embasadas nessas funções estimáveis, mas sim que, vários usuários não iniciados nos princípios dos testes de hipóteses, nem imaginam o que estão testando.

A hipótese sobre a interação é sempre ajustada para os demais fatores, portanto todos os tipos de somas de quadrados se equivalem e são iguais a $R(\gamma|\mu,\alpha,\beta)$, fato que vem confirmar a idéia de Iemma (1995), de que modelo com interação é próprio para testar interação.

Em se tratando de modelo com interação com dados desbalanceados e caselas vazias, as quatro somas de quadrados fornecidas pelo PROC GLM do SAS diferem entre si e as somas de quadrados Tipo I dependem da ordem de entrada dos parâmetros no modelo, dados que são obtidos seqüencialmente. Na Tabela 4, mostra-se que $R(\alpha|\mu) \neq R(\alpha|\mu,\beta)$ e $R(\beta|\mu) \neq R(\beta|\mu,\alpha)$. Portanto, se a ordem de entrada dos parâmetros é A, B, A-B, ou seja, o modelo é $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$, as somas de quadrados são obtidas primeiramente supondo-se o modelo $y_{ijk} = \mu + \alpha_i + e_{ijk}$, seguido do modelo $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$, e pelo modelo completo, gerando respectivamente $R(\alpha|\mu) = 1542{,}597$, $R(\beta|\mu,\alpha) = 1270{,}737$ e $R(\gamma|\mu,\alpha,\beta) = 175{,}559$. Semelhante, se a ordem de entrada é B, A, B-A, as somas de quadrados são obtidas por meio dos modelos $y_{jik} = \mu + \beta_j + e_{jik}$, $y_{jik} = \mu + \beta_j + \alpha_i + e_{jik}$ e por fim o modelo completo $y_{jik} = \mu + \beta_j + \alpha_i + \gamma_{ij} + e_{jik}$, obtendo-se respectivamente $R(\beta|\mu) = 1424{,}937$, $R(\alpha|\mu,\beta) = 1388{,}370$ e $R(\gamma|\mu,\alpha,\beta) = 175{,}559$. De acordo com Camarinha Filho (1995), as somas de quadrados do Tipo I e II, testam a mesma hipótese para o segundo fator na ordenação do modelo.

A ordem dos fatores no modelo não altera os resultados das somas de quadrados tipo III e IV. Resultados semelhantes foram verificados por Iemma (1995) em fatorial 2 x 3, com dados desbalanceados e uma casela vazia e por Nekatschalow (1997), em um fatorial 3 x 4 com duas caselas vazias.

A soma de quadrado Tipo IV não é única, o que causa uma maior dificuldade de interpretação, pois com uma reordenação dos dados há possibilidade de obter outras funções estimáveis do Tipo IV, induzindo a outras somas de quadrados.

A ocorrência de dados desbalanceados em presença de caselas vazias pode trazer certos transtornos aos usuários das ciências aplicadas, com relação à identificação das hipóteses estatísticas. Assim sendo, usuários comuns de programas estatísticos devem ser cautelosos, evitando o uso indiscriminado desses programas sem conhecimento adequado de suas documentações.

TABELA 4 – Análise de variância para os dados da Tabela 2, fornecida pelo PROC GLM do SAS,com base no modelo $y_{jik} = \mu + \beta_j + \alpha_i + \gamma_{ij} + e_{jik}$.

Ordem A-B					Ordem B-A				
C.V	G.L	H_0	R( )	SS	C.V	G.L	H_0	R( )	SS
TIPO I									
A	3	H_0^1	R(α|μ)	1542,597	B	3	H_0^5	R(β|μ)	1424,937
B(aj)	3	H_0^6	R(β|μ,α)	1270,737	A(aj)	3	H_0^2	R(α|μ,β)	1388,370
A-B	7	H_0^9	R(γ|μ,α,β)	175,559	B-A	9	H_0^9	R(γ|μ,α,β)	175,559
TIPO II									
A(aj)	3	H_0^2	R(α|μ,β)	1388,370	B(aj)	3	H_0^6	R(β|μ,α)	1270,737
B(aj)	3	H_0^6	R(β|μ,α)	1270,737	A(aj)	3	H_0^2	R(α|μ,β)	1388,370
A-B	9	H_0^9	R(γ|μ,α,β)	175,559	B-A	9	H_0^9	R(γ|μ,α,β)	175,559
TIPO III									
A	3	H_0^3	R(α̇|μ̇,β̇,γ̇)	1379,424	B	3	H_0^7	R(β̇|μ̇,α̇,γ̇)	1364,799
B	3	H_0^7	R(β̇|μ̇,α̇,γ̇)	1264,799	A	3	H_0^3	R(α̇|μ̇,β̇,γ̇)	1379,424
A-B	9	H_0^9	R(γ̇|μ̇,α̇,β̇)	175,559	B-A	9	H_0^9	R(γ̇|μ̇,α̇,β̇)	175,559
TIPO IV									
A	3(1)	H_0^4	—	1162,994	B	3(1)	H_0^4	—	880,698
B	3(1)	H_0^8	—	880,698	A	3(1)	H_0^8	—	1162,994
A-B	9	H_0^9	—	175,559	B-A	9	H_0^9	—	175,559

(1) Existem outras hipóteses testáveis Tipo IV que produzem diferentes somas de quadrados.

CONCLUSÕES

a) Quando os dados são desbalanceados, a identificação de funções paramétricas estimáveis é complexa e as hipóteses são de difícil interpretação.

b) Quando os dados são desbalanceados em relação ao número de repetições:

- as funções estimáveis de um fator envolvem os parâmetros relativos ao fator e os componentes das interações nas quais o fator está presente;
- as somas de quadrados dos Tipos III e IV são iguais;
- a ordenação dos fatores principais afeta as hipóteses do Tipo I.

c)Quando os dados são desbalanceados em relação às combinações dos fatores:

- os quatro tipos de somas de quadrados para o fator principal de entrada, foram diferentes;
- a ordenação é fundamental para obtenção das hipóteses do Tipo I.

d) Um pesquisador ao analisar dados desbalanceados, em modelos com interação e na presença de caselas vazias, deve ter o cuidado de verificar as hipóteses a serem testadas.

REFERÊNCIAS BIBLIOGRÁFICAS

CAMARINHA FILHO, J. A. **Testes de hipóteses em modelos lineares com dados desbalanceados e caselas vazias**. 1995. 142 f. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1995.

FREUND, R. J. The case of the missing cell. **The American statistician**, Alexandria, v. 34, n. 2, p. 94-98, May 1980.

HERR, D. G. On the history of ANOVA in unbalanced, factorial designs: the first 30 years. **The American Statistician**, Alexandria, v. 40, n. 4, p. 265-270, Nov. 1986.

IEMMA, A. F. Que hipóteses estatísticas testamos através do SAS em presença de caselas vazias? **Scientia Agrícola**, Piracicaba, v. 52, n. 2, p. 210-220, maio/jun. 1995.

MONDARDO, M. **Estimabilidade de funções paramétricas com dados desbalanceados através do PROC GLM do SAS**: aplicações à pesquisa agropecuária. 1994. 166 f. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1994.

NEKATSCHALOW, M. C. **Análise de variância**: alternativas através de modelos de posto completo. 1997. 123 f. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1997.

SANTOS, E. S. **Testes de hipóteses com dados desbalanceados e interpretação de *softwares* mais utilizados**. 1994. 230 f. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1994.

SAS INSTITUTE. **User's guide**: statistics. Version 6. Cary, 1990. 846 p.

SEARLE, S. R. **Linear models**. New York: J. Wiley & Sons, 1971. 532 p.

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

SEARLE, S. R.; SPEED, F. M.; MILLIKEN, G. A. Population marginal means in the linear model: an alternative to least squares means. **The American Statistician**, Alexandria, v. 34, n. 4, p. 216-221, Nov. 1980.

SPEED, F. M.; HOCKING, R. R. The use of the R( )-notation with unbalanced data. **The American Statistician**, Washington, v. 28, n. 1, p. 30-33, Feb. 1976.

SPEED, F. M.; HOCKING, R. R.; HACKNEY, O. P. Methods of analysis of linear models with unbalanced data. **Journal of the American Statistical Association**, Boston, v. 73, n. 361, p. 105-112, Mar. 1978.

WECHSLER, F. S. Fatoriais fixos desbalanceados: uma análise mal compreendida. **Pesquisa Agropecuária Brasileira**, Brasília, v. 33, n. 3, p. 231-262, mar. 1998.

YATES, F. The analysis of multiple classifications with unequal numbers in the different classes. **Journal of the American Statistical Association**, Alexandria, v. 29, n. 1, p. 51-66, 1934.

YATES, F. The principles of orthogonality and confounding in replicated experiments. **Journal Agricultural Science**, Cambridge, v. 23, pt. 1, p. 108-145, Jan. 1933.