USO DE *SOFTWARES* ESTATÍSTICOS NA ANÁLISE DE VARIÂNCIA DE DADOS DESBALANCEADOS EM MODELOS COM DOIS FATORES CRUZADOS DE EFEITOS FIXOS

Eduardo Yoshio Nakano
Universidade de Brasília - UnB
Departamento de Estatística,
Campus Darcy Ribeiro, 70910-900, Brasília/DF.
nakano@unb.br

Sérgio Minoru Oikawa
Universidade Estadual Paulista - UNESP
Departamento de Matemática, Estatística e Computação
Faculdade de Ciências e Tecnologia - FCT
Rua Roberto Simonsen 305, 19060-900, Presidente Prudente/SP.
smoikawa@prudente.unesp.br

Resumo. A interpretação das hipóteses estatísticas testadas através da análise de variância de dados balanceados pode ser feita, em geral, sem dificuldade. No entanto, a ocorrência de desbalanceamento nos dados pode gerar equívocos em sua interpretação. A falta de informação sobre quais hipóteses estão sendo testadas por um determinado software pode induzir o pesquisador ao erro, comprometendo o resultado das pesquisas. Este trabalho teve como objetivo comparar os resultados de outros softwares estatísticos com aqueles fornecidos pelo procedimento GLM do SAS. O problema de interpretação das hipóteses testadas nos modelos com dois fatores cruzados e dados desbalanceados foi amplamente discutido com base no procedimento GLM do SAS. Concluímos que os usuários de softwares estatísticos devem ser cautelosos na análise estatística de dados desbalanceados, evitando o uso indiscriminado dos softwares sem o conhecimento prévio de sua documentação.

Palavras-chave: dados desbalanceados, caselas vazias, hipóteses estatísticas, softwares, somas de quadrados.

Abstract. The interpretation of the statistical hypothesis tested in the analysis of variance of balanced data usually can be performed with no difficulty. However, the presence of unbalanced data can result in errors of interpretation. The lack of information about the hypothesis being tested in a software can induce the researcher to a mistake, compromising the results of the study. The purpose of the present work was to compare the results of others softwares with those provided by GLM procedure of the SAS. The interpretation problem of tested hypothesis in the models with two crossed factors and unbalanced data was widely argued on the basis of GLM procedure of the SAS. We concluded that the users of statistical softwares must be very cautious in the analysis of unbalanced data, and that they need to have a previous knowledge of the software before using it.

Keywords: unbalanced data, empty cells, statistics hypothesis, softwares, sum of squares.

1 INTRODUÇÃO

Atualmente, os *softwares* estatísticos tornaram-se uma ferramenta importante e indispensável na análise estatística de dados. Segundo Searle [4], os *softwares* estatísticos, hoje disponíveis, são capazes de realizar cálculos aritméticos complexos que eram totalmente inconcebíveis há algum tempo atrás. As capacidades dos computadores de hoje, sua grande rapidez e seu baixo custo operacional por unidade aritmética, eram características totalmente inimagináveis para muitos estatísticos da época. Tais características marcantes, bem como a facilidade de acesso, fizeram com que o número de usuários crescesse consideravelmente. Segundo Iemma [1], esse fato acabou gerando um sério problema motivado pela utilização muitas vezes inadequada dos *softwares* estatísticos.

Neste contexto, é importante chamar a atenção para as "interpretações das verdadeiras hipóteses" testadas através das somas de quadrados obtidas pelos diversos métodos disponíveis na literatura. Por exemplo, o procedimento GLM (General Linear Models) do SAS (Statistical Analysis System) fornece quatro tipos de somas de quadrados (I, II, III e IV) que, dependendo do nível de desbalanceamento e da posição das caselas vazias, testam quatro tipos diferentes de hipóteses.

Visando exemplificar numericamente os conceitos sobre dados desbalanceados e caselas vazias, apresenta-se aqui, parte dos dados provenientes do peso de bezerros da raça Canchim e reproduzidos em Oikawa [3]. Os dados apresentados seguem a estrutura de um modelo com dois fatores, em que as fontes de variações (fatores) são dadas pelo sexo do bezerro (efeito de linhas) e pela origem do bezerro (efeito de colunas). O Quadro 1 apresenta os dados desbalanceados com todas caselas ocupadas, ou seja, o número de repetições não é o mesmo em cada casela, no entanto todas caselas apresentam, pelo menos, uma observação. No Quadro 2, os dados estão desbalanceados e existe a ocorrência de uma casela vazia.

Quadro 1. Peso a desmama, em kg, de bezerros da raça Canchim.

	Touro 1 (B_1)	Touro 2 (B_2)	Touro 3 (B_3)
Macho (A_1)	120; 152	167; 172	209
Fêmea (A_2)	157; 150; 160; 130	185; 153; 173; 191; 160; 224	169; 187; 224

Fonte: Centro de Pesquisa de Pecuária do Sudeste – CPPSE/EMBRAPA, São Carlos – SP.

Quadro 2. Dados adaptados do Quadro 1 com o objetivo de obter casela vazia.

	Touro 1 (B1)	Touro 2 (B_2)	Touro 3 (B_3)
Macho (A_1)	120; 152	167; 172	
Fêmea (A_2)	157; 150; 160; 130	185; 153; 173; 191; 160; 224	169; 187; 224

Fonte: Centro de Pesquisa de Pecuária do Sudeste – CPPSE/EMBRAPA, São Carlos – SP.

Se os dados são balanceados (todas as caselas têm o mesmo número de repetições), não existem dificuldades para as interpretações das hipóteses testadas, pois elas são todas equivalentes. Neste caso, os usuários, especialmente os pesquisadores das ciências aplicadas, podem interpretar facilmente as suas hipóteses que são testadas pelos diversos *softwares* estatísticos existentes. Logo, as análises estatísticas podem ser realizadas através do *software* de sua preferência ou disponibilidade.

No entanto, se os dados são desbalanceados com presença ou não de caselas vazias (Quadro 1 ou 2), os métodos disponíveis fornecem diferentes somas de quadrados e, portanto, testam diferentes hipóteses. Desse modo, os pesquisadores não iniciados na análise estatística de dados desbalanceados podem estar testando hipóteses completamente diferentes daquelas que eles julgam testar, alterando sensivelmente as interpretações de seus resultados. Existem, na literatura, inúmeros artigos que tratam sobre o assunto. Porém, em casos de dados desbalanceados com caselas vazias, ainda há muitas controvérsias sobre a utilização de uma metodologia unificada. Evidentemente, isso reflete sobre os *softwares* estatísticos que utilizam diferentes métodos, fornecendo diferentes resultados para o mesmo conjunto de dados.

Vários autores discutem o problema da interpretação das hipóteses testadas com base no procedimento GLM do SAS. A escolha desse *software* pode ter sido justificada pelo fato do SAS fornecer, além das somas de quadrados, os quatro tipos de funções estimáveis. A importância das funções estimáveis está no reconhecimento da hipótese testada por uma determinada soma de quadrados (Mondardo, [2]).

Neste contexto, este trabalho teve como objetivo comparar, sem o apelo de competição, os resultados apresentados por outros *softwares* estatísticos com aqueles fornecidos pelo procedimento GLM do SAS. Para tanto, foram utilizados os seguintes *softwares*: BMDP, MINITAB, NTIA, SPSS, STATISTICA e S-PLUS.

2 DESENVOLVIMENTO

Para ilustrar os procedimentos descritos, foram utilizados os dados apresentados nos Quadros 1 e 2. Com base neste conjunto de dados, discutiram-se os resultados fornecidos pelo procedimento GLM do SAS, visando interpretar as hipóteses estatísticas e somas de quadrados a elas associadas. Para tanto, considerou-se o modelo com dois fatores cruzados de efeitos fixos:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (1)$$

onde, y_{ijk} é a k-ésima observação na linha i e coluna j ; μ é a média geral ; α_i é o efeito devido a i-ésima linha; β_j é o efeito devido a j-ésima coluna ; γ_{ij} é a interação entre os efeitos da i-ésima linha com a j-ésima coluna e ε_{ijk} são variáveis aleatórias não observáveis, tais que, $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Dentre os vários tipos de hipóteses existentes, o procedimento GLM do SAS incorporou, em relação ao modelo em estudo, quatro tipos de somas de quadrados para efeitos de linhas, quatro para efeitos de colunas e um para a interação (fator cruzado).

Neste contexto, visando simplificar a interpretação do leitor, as hipóteses de interesse serão rotuladas segundo as somas de quadrados a elas associadas. Neste trabalho, as somas de quadrados serão representadas através da notação R(.) (Searle, [4]). Assim, têm-se as seguintes hipóteses de interesse:

Hipóteses mais comuns sobre efeitos de linhas:

$H_0^{(1)}$: "hipóteses sobre médias ponderadas de linhas não ajustadas"

$S.Q.H_0^{(1)} = R(\alpha|\mu)$

$H_0^{(2)}$: "hipóteses sobre médias ponderadas de linhas ajustadas para colunas"

$S.Q.H_0^{(2)} = R(\alpha|\mu, \beta)$

$H_0^{(3)}$: "hipóteses sobre médias não ponderadas de linhas"

$S.Q.H_0^{(3)} = R(\dot{\alpha} \mid \dot{\mu}, \dot{\beta}, \dot{\gamma})$

Hipóteses mais comuns sobre efeitos de colunas:

$H_0^{(4)}$: "hipóteses sobre médias ponderadas de colunas não ajustadas"

$S.Q.H_0^{(4)} = R(\beta|\mu)$

$H_0^{(5)}$: "hipóteses sobre médias ponderadas de colunas ajustadas para linhas"

$S.Q.H_0^{(5)} = R(\beta|\mu, \alpha)$

$H_0^{(6)}$: "hipóteses sobre médias não ponderadas de colunas"

$S.Q.H_0^{(6)} = R(\dot{\beta} \mid \dot{\mu}, \dot{\alpha}, \dot{\gamma})$

Hipótese mais comum sobre a interação:

$H_0^{(7)}$: "hipótese sobre a interação"

$S.Q.H_0^{(7)} = R(\gamma|\mu, \alpha, \beta)$

2.1 Dados desbalanceados com todas caselas ocupadas

Os comandos utilizados pelo procedimento GLM do SAS para o cálculo das somas de quadrados (dos dados do Quadro 1) são dados por:

```
DATA EXEMPLO;
INPUT A B Y;
CARDS;
1 1 120
1 1 152
1 2 167
⋮
2 3 224
;
PROC GLM;
CLASS A B;
MODEL Y = A B A*B / SS1 SS2 SS3 SS4;
RUN;
```

O Quadro 3 apresenta os resultados da análise de variância fornecido pelo procedimento GLM do SAS, considerando o conjunto de dados do Quadro 1.

Quadro 3. *Somas de quadrados (Tipo I, II, III e IV) fornecidos pelo SAS-GLM a partir dos dados do Quadro 1, utilizando o modelo (1) e seguindo a ordenação A, B, AB.*

S.Q.Tipo I				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo I**
A (não ajustado)	1	$H_0^{(1)}$	$R(\alpha \mid \mu)$	240,0855
B (ajustado)	2	$H_0^{(5)}$	$R(\beta \mid \mu, \alpha)$	6809,9714
AB	2	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	418,9709
S.Q.Tipo II				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo II**
A (ajustado)	1	$H_0^{(2)}$	$R(\alpha \mid \mu, \beta)$	79,8624
B (ajustado)	2	$H_0^{(5)}$	$R(\beta \mid \mu, \alpha)$	6809,9714
AB	2	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	418,9709
S.Q.Tipo III e IV				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo III e IV**
A	1	$H_0^{(3)}$	$R(\dot{\alpha}\|\dot{\mu}, \dot{\beta}, \dot{\gamma})$	8,7904

B	2	$H_0^{(6)}$	$R(\dot{\beta}\|\dot{\mu},\dot{\alpha},\dot{\gamma})$	6892,7035
AB	2	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	418,9709

2.2 Dados desbalanceados com caselas vazias

O Quadro 4 apresenta os resultados da análise de variância fornecidos pelo procedimento GLM do SAS, considerando o conjunto de dados do Quadro 2.

Quadro 4. Somas de quadrados (Tipo I, II, III e IV) fornecidos pelo SAS-GLM a partir dos dados do Quadro 2, utilizando o modelo (1) e a ordenação A, B, AB.

S.Q.Tipo I				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo I**
A (não ajustado)	1	$H_0^{(1)}$	$R(\alpha \mid \mu)$	1151,6753
B (ajustado)	2	$H_0^{(5)}$	$R(\beta \mid \mu, \alpha)$	4672,9815
AB	1	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	24,7108
S.Q.Tipo II				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo II**
A (ajustado)	1	$H_0^{(2)}$	$R(\alpha \mid \mu, \beta)$	290,0392
B (ajustado)	2	$H_0^{(5)}$	$R(\beta \mid \mu, \alpha)$	4672,9815
AB	1	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	24,7108
S.Q.Tipo III				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo III**
A	1	$H_0^{(3)}$	$R(\dot{\alpha}\|\dot{\mu},\dot{\beta},\dot{\gamma})$	299,0637
B	2	$H_0^{(6)}$	$R(\dot{\beta}\|\dot{\mu},\dot{\alpha},\dot{\gamma})$	4494,9157
AB	1	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	24,7108
S.Q.Tipo IV				
Variações Consideradas	**Graus de Liberdade**	**Hipóteses Testadas**	**R(.)**	**S.Q. Tipo IV**
A	1*	$H_0^{(3)}$*	$R(\dot{\alpha}\|\dot{\mu},\dot{\beta},\dot{\gamma})$	299,0637
B	2*	$H_0^{(8)}$*	S.Q. $H_0^{(8)}$	3575,4423
AB	1	$H_0^{(7)}$	$R(\gamma \mid \mu, \alpha, \beta)$	24,7108

*Notas: *Existem outras hipóteses do Tipo IV associadas aos efeitos A e B, podendo, assim, resultar em diferentes somas de quadrados.*

As somas de quadrados do Tipo I fornecidas pelo procedimento GLM do SAS são obtidas seqüencialmente, ou seja, as hipóteses são montadas ajustando um parâmetro após o outro. Assim, apenas o primeiro parâmetro do quadro de análise de variância está associado à hipótese sobre as médias ponderadas não ajustadas. Como se pôde observar nos Quadros 3 e 4, a soma de quadrados do Tipo I não testa a hipótese de médias ponderadas de colunas não ajustadas, $H_0^{(4)}$, visto que o modelo segue a ordenação A, B e AB. Portanto, a ordem de entrada dos fatores no modelo é de fundamental importância para a obtenção das hipóteses sobre as médias ponderadas não ajustadas. Desta forma, para testar a hipótese $H_0^{(4)}$ deve-se considerar o modelo com a ordenação B, A, AB. Já as somas de quadrados do Tipo II testam as hipóteses sobre as médias ponderadas ajustadas. As somas de quadrados do Tipo III testam as hipóteses sobre as médias não ponderadas. É importante ressaltar que, quando há caselas vazias, a soma de quadrados do Tipo III estará testando as hipóteses sobre médias ponderadas de linhas (colunas) "nas colunas (linhas) completas".

Como visto no Quadro 3, as somas de quadrados do Tipo IV são similares as do Tipo III se não existem caselas vazias. Se, no entanto, existe ao menos uma casela vazia, então as somas de quadrados dos Tipos III e IV são, em geral, diferentes e podem não ser únicas, pois elas dependem da posição e do número de caselas vazias.

De modo geral, as somas de quadrados do Tipo IV testam as hipóteses sobre contrastes entre médias das caselas que estão na mesma coluna (linha). O pesquisador deve, então, tomar cuidado ao interpretar a hipótese

testada quando existem caselas vazias, pois neste caso a hipótese correspondente, em geral, não considera todos os parâmetros

Se os dados são balanceados, então as somas de quadrados do Tipo I, II, III e IV são equivalentes e testam sempre a mesma hipótese.

3 HIPÓTESES TESTADAS POR OUTROS *SOFTWARES*

Apresentamos aqui uma comparação, sem o apelo de competição, das hipóteses testadas através do procedimento GLM do SAS com aquelas testadas por outros *softwares* estatísticos. Para tanto, foram utilizados os *softwares*: BMDP, MINITAB, NTIA, SPSS, STATISTICA e S-PLUS. Ressalta-se aqui que os *softwares* estatísticos são abordados do ponto de vista do usuário e não do ponto de vista do especialista. Neste contexto, são utilizados apenas comandos básicos usuais e não programações mais sofisticadas.

3.1 Dados desbalanceados com todas caselas ocupadas

O Quadro 5 apresenta as hipóteses do modelo (1) testadas por diversos *softwares* estatísticos, considerando o conjunto de dados do Quadro 1.

Quadro 5. Análise de variância do modelo (1) fornecida por diversos softwares, a partir dos dados do Quadro 1.

SOFTWARE	Variações Consideradas	Graus de Liberdade	Hipóteses Testadas	
			Ordenação A, B, AB	Ordenação B, A, AB
MINITAB NTIA SPLUS	A	1	$H0^{(1)}$, $H0^{(3)}$	$H0^{(2)}$, $H0^{(3)}$
	B	2	$H0^{(5)}$, $H0^{(6)}$	$H0^{(4)}$, $H0^{(6)}$
	AB	2	$H0^{(7)}$	$H0^{(7)}$
STATISTICA	A	1	$H0^{(1)}$, $H0^{(2)}$, $H0^{(3)}$	$H0^{(1)}$, $H0^{(2)}$, $H0^{(3)}$
	B	2	$H0^{(4)}$, $H0^{(5)}$, $H0^{(6)}$	$H0^{(4)}$, $H0^{(5)}$, $H0^{(6)}$
BMDP	AB	2	$H0^{(7)}$	$H0^{(7)}$
SPSS	A	1	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(2)}$, $H_0^{(3)}$
	B	2	$H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$
	AB	2	$H_0^{(7)}$	$H_0^{(7)}$

3.2 Dados desbalanceados com caselas vazias

O Quadro 6 apresenta as hipóteses do modelo (1) testadas por diversos *softwares* estatísticos, considerando o conjunto de dados do Quadro 2.

Quadro 6. Análise de variância do modelo (1) fornecida por diversos softwares, a partir dos dados do Quadro 2.

SOFTWARE	Variações Consideradas	Graus de Liberdade	Hipóteses Testadas	
			Ordenação A, B, AB	Ordenação B, A, AB
MINITAB	A	1	$H_0^{(1)}$	$H_0^{(2)}$
	B	2	$H_0^{(5)}$	$H_0^{(4)}$
NTIA	AB	1	$H_0^{(7)}$	$H_0^{(7)}$
BMDP	A	1	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$
	B	2	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$
	AB	1	$H_0^{(7)}$	$H_0^{(7)}$
SPSS	A	1	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(2)}$, $H_0^{(3)}$
	B	2	$H_0^{(5)}$, $H_0^{(6)}$, $H_0^{(8)}$	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$, $H_0^{(8)}$
	AB	1	$H_0^{(7)}$	$H_0^{(7)}$
S-PLUS	A	1	$H0^{(1)}$, $H0^{(3)}$	$H0^{(2)}$, $H0^{(3)}$
	B	2	$H0^{(5)}$, $H0^{(6)}$	$H0^{(4)}$, $H0^{(6)}$
	AB	1	$H0^{(7)}$	$H0^{(7)}$
STATISTICA	O STATISTICA não fornece as somas de quadrados se existem caselas vazias			

4 DISCUSSÕES

4.1 MINITAB – versão12.23

As somas de quadrados foram obtidas através do procedimento GLM do MINITAB através dos seguintes comandos:

```
MTB> NAME C1= 'A' C2='B' C3='Y'
MTB> GLM Y=A|B
```

Em geral, quando o modelo utilizado é com interação e todas as caselas estão ocupadas, o procedimento GLM do MINITAB fornece somas de quadrados dos tipos seqüenciais e ajustadas, equivalentes às somas de quadrados dos Tipos I e II do SAS.

Se há caselas vazias, o MINITAB fornece apenas as somas de quadrados do tipo seqüencial, isto é, as somas de quadrados do Tipo I fornecidas pelo SAS. Ademais, o MINTAB emite a seguinte mensagem: "+ *Rank deficiency due to empty cells, unbalanced nesting or collinearity. No storange of results or further analysis will be done.*", e então não fornece as somas de quadrados ajustadas.

4.2 NTIA – versão 4.2.2

Os comandos para o cálculo das somas de quadrados pelo *software* NTIA são:

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

Em casos onde todas as caselas estão ocupadas, o *software* NTIA tem o mesmo desempenho do MINITAB, fornecendo as somas de quadrados dos tipos seqüenciais e ajustadas (Tipo I e III do SAS).

Quando há presença de casela vazia, o NTIA falha em fornecer as somas de quadrados do Tipo III. Ademais, o usuário é alertado do motivo pelo qual as somas de quadrados parciais não são fornecidas.

4.3 STATISTICA – versão 5.0

Os cálculos das somas de quadrados foram realizados pelo módulo "General ANOVA/MANOVA" do *software* STATISTICA.

No caso em que todas as caselas estão ocupadas, o STATISTICA fornece três tipos de somas de quadrados: Tipo I, II e III. E, ao contrário do SAS, o STATISTICA não depende da ordenação do modelo para fornecer suas somas de quadrados. As somas de quadrados do Tipo I são equivalentes às somas de quadrados não ajustadas, ou seja, são referentes às hipóteses sobre médias ponderadas não ajustadas. As somas de quadrados do Tipo II correspondem às somas de quadrados ajustadas e testam as hipóteses sobre médias ponderadas ajustadas. As somas de quadrado do Tipo II testam as hipóteses sobre médias não ponderadas.

Em caso de casela vazia, o STATISTICA emite a mensagem "DESIGN INCOMPLETE ; REGRESSION APPROACH NOT AVAILABLE" e não calcula as somas de quadrados Tipo I, II e III, como faz o procedimento GLM do SAS.

4.4 BMDP – versão PC90

Os comandos do *software* BMDP para os cálculos das somas de quadrados (do Quadro 1) são dados por:

```
/ INPUT TITLE IS 'EXEMPLO_OCUP'.
        VARIABLES = 3.
        FORMAT = FREE.
/ VARIABLE      NAMES = A, B, Y.
/ BETWEEN       FACTORS = A, B.
 CODES(A) = 1, 2.
 CODES(B) = 1 TO 3.
/ WEIGHTS       BTWEEN = EQUAL.
/ END
1 1 120
⋮
2 3 224
/ END
ANALYSIS        PROCEDURE               =
STRUCTURE. /
        BFORMULA = 'A*B'. /
END /
/ WEIGHT        BETWEEN = SIZES.
/ END
ANALYSIS        PROCEDURE               =
STRUCTURE. /
        BFORMULA = 'A*B'. /
END /
```

O BMDP possui dois comandos para realizar a análise de variância: "BETWEEN = SIZES" e "BETWEEN = EQUAL". Da mesma forma que o STATISTICA, o BMDP não depende da ordenação do modelo para fornecer suas somas de quadrados. Ademais, tem-se que, tanto no caso onde todas as caselas estão ocupadas como no caso em que há caselas vazias, o comando "BETWEEN = SIZES" fornece somas de quadrados não ajustadas e ajustadas equivalentes às somas de quadrados dos Tipos I e II do SAS e o comando "BETWEEN = EQUAL" fornece as somas de quadrados parciais, equivalentes às somas de quadrados do Tipo III do SAS. Portanto, o BMDP apresenta a mesma performance independentemente se há ou não caselas vazias.

4.5 S-PLUS – versão 2.000

Tanto no caso onde todas as caselas estão ocupadas como no caso onde há caselas vazias, o módulo "ANOVA/Fixed Effects" do S-PLUS fornece as somas de quadrados dos Tipos I e III, equivalentes às somas de quadrados seqüenciais e parciais, respectivamente.

4.6 SPSS – versão 10.0

Tanto no caso onde todas as caselas estão ocupadas como no caso onde há caselas vazias, o procedimento "GENERAL LINEAR MODELS - UNIVARIATE" do SPSS fornece as mesmas somas de quadrados do procedimento GLM do SAS, sem nenhuma exceção. O SPSS calcula todas as somas de quadrados fornecidas pelo SAS (Tipo I, II, III e IV), sendo que as somas de quadrados do Tipo III é o seu *default*.

5 CONCLUSÕES

Com base nas discussões, verificou-se que a ocorrência de desbalanceamento nos dados pode trazer sérios transtornos aos pesquisadores das ciências aplicadas, pois, na maioria dos casos, a falta de uma documentação explícita sobre quais hipóteses esses *softwares* estão testando pode induzir a tomadas de decisões incorretas, comprometendo o resultado de suas pesquisas.

Sendo assim, os usuários de *softwares* estatísticos devem ser cautelosos na análise estatística de dados desbalanceados, evitando o uso indiscriminado dos *softwares* sem o conhecimento prévio de sua documentação.

6 AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro concedido para o desenvolvimento deste trabalho. Processo nº 112655/1999-8.

REFERÊNCIAS

[1] IEMMA. A. F. *Análise de variância de dados desbalanceados*. 4º Congresso Brasileiro de Usuários do SAS. Universidade de São Paulo. 1995, 111p.

[2] MONDARDO, M. *Estimabilidade de funções paramétricas com dados desbalanceados através do PROC-GLM do SAS: Aplicações à pesquisa agropecuária*. Piracicaba, 1994, 166 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica), ESALQ-USP.

[3] OIKAWA, S. M. *Hipóteses estatísticas com dados desbalanceados nos modelos de efeitos fixos hierarquizados em presença ou não de esquema fatorial*. Piracicaba, 1998, 145p. Tese (Doutorado em Estatística e Experimentação Agronômica), ESALQ-USP.

[4] SEARLE, S. R. *Linear Models for Unbalanced Data*. New York: John Wiley & Sons, 1987, 536p.