TEMA Tend. Mat. Apl. Comput., **5**, No. 1 (2004), 117-124.

Hipóteses Estatísticas Testadas por Diversos *Softwares* em Modelos com Dois Fatores Cruzados e Dados Desbalanceados

E.Y. NAKANO, Departamento de Estatística, Universidade de Brasília, UnB, 70910-900 Brasília, DF, Brasil

S.M. OIKAWA, Departamento de Matemática, FCT, UNESP, 19060-900 Presidente Prudente, SP, Brasil.

Resumo. 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 (Iemma [1] e Mondardo [2]). Neste contexto, este trabalho teve como objetivo comparar os resultados de outros *softwares* estatísticos com aqueles fornecidos pelo procedimento GLM do SAS.

1. Introdução

Devido a capacidade dos computadores de hoje, sua grande rapidez na realização de cálculos aritméticos e facilidade de acesso, os *softwares* estatísticos tornaram-se uma ferramenta indispensável na análise estatística de dados. A utilização muitas vezes inadequada desses *softwares* acabou gerando um dos problemas mais comuns em estatística, que é a interpretação das hipóteses testadas na análise de variância de dados desbalanceados.

Segundo Iemma [1], 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, onde as fontes de variações (fatores) são dadas pelo sexo do bezerro (efeito de linhas) e pela origem do bezerro (efeito de colunas). O Quadro I 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 II os dados estão desbalanceados e existe a ocorrência de uma casela vazia.

Quadro I: 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 II: Dados adaptados do Quadro I com o objetivo de obter casela vazia.

	Touro 1 (B_1)	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. 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 I ou II), os métodos disponíveis fornecem diferentes somas de quadrados e, portanto, testam diferentes hipóteses. 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 I e II. Com base neste conjunto de dados, discutiu-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:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{2.1}$$

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 $R(.)$, muito utilizada por 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"

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

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"

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

Hipótese mais comum sobre a interação:

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

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

Os Quadros III e IV apresentam os quatro tipos de somas de quadrados fornecidos pelo procedimento GLM do SAS. A análise foi realizada utilizando o modelo (2.1) com a ordenação A, B, AB. O Quadro III apresenta as somas de quadrados obtidas considerando o conjunto de dados do Quadro I (dados desbalanceados com todas caselas ocupadas) e o Quadro IV apresenta as somas de quadrados obtidas a partir do conjunto de dados do Quadro II (dados desbalanceados com caselas vazias).

Quadro III: Somas de quadrados fornecidos pelo procedimento GLM do SAS a partir dos dados do Quadro I, considerando a ordenação A, B, AB.

Tipo I				
variações consideradas	G.L.	hipóteses testadas	$R(.)$	S.Q. Tipo I
A (não ajustado)	1	$H_0^{(1)}$	$R(\alpha\|\mu)$	$240,0855$
B (ajustado)	2	$H_0^{(5)}$	$R(\beta\|\mu,\alpha)$	$6809,9714$
AB	2	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	$418,9709$
Tipo II				
variações consideradas	G.L.	hipóteses testadas	$R(.)$	S.Q. Tipo II
A (ajustado)	1	$H_0^{(2)}$	$R(\alpha\|\mu,\beta)$	$79,8624$
B (ajustado)	2	$H_0^{(5)}$	$R(\beta\|\mu,\alpha)$	$6809,9714$
AB	2	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	$418,9709$
Tipo III e IV				
variações consideradas	G.L.	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	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	$418,9709$

As somas de quadrados do Tipo I fornecidas pelo procedimento GLM do SAS são obtidas seqüêncialmente, 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 pode observar nos Quadros III e IV, 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 III, 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).

Quadro IV: Somas de quadrados fornecidos pelo procedimento GLM do SAS a partir dos dados do Quadro II, considerando a ordenação A, B, AB.

Tipo I				
variações consideradas	G.L.	hipóteses testadas	$R(.)$	S.Q. Tipo I
A (não ajustado)	1	$H_0^{(1)}$	$R(\alpha\|\mu)$	1151,6753
B (ajustado)	2	$H_0^{(5)}$	$R(\beta\|\mu,\alpha)$	4672,9815
AB	1	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	24,7108
Tipo II				
variações consideradas	G.L.	hipóteses testadas	$R(.)$	S.Q. Tipo II
A (ajustado)	1	$H_0^{(2)}$	$R(\alpha\|\mu,\beta)$	290,0392
B (ajustado)	2	$H_0^{(5)}$	$R(\beta\|\mu,\alpha)$	4672,9815
AB	1	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	24,7108
Tipo III				
variações consideradas	G.L.	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	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	24,7108
Tipo IV				
variações consideradas	G.L.	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	$H_0^{(7)}$	$R(\gamma\|\mu,\alpha,\beta)$	24,7108

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

** Hipótese sobre contrastes entre médias das caselas que estão na mesma coluna.

3. Hipóteses Testadas por Outros *Softwares*

Apresenta-se 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.

Os Quadros V e VI apresentam as hipóteses do modelo (2.1) testadas por diversos *softwares* estatísticos. O Quadro V apresenta os resultados obtidos considerando o conjunto de dados do Quadro I (dados desbalanceados com todas caselas ocupadas) e o Quadro VI apresenta os resultados obtidos a partir do conjunto de dados do Quadro II (dados desbalanceados com caselas vazias).

Quadro V: Análise de variância do modelo (2.1) fornecida por diversos softwares, a partir dos dados do Quadro I.

Software	Variações	Hipóteses Testadas	
	Consideradas	Ordenação A, B, AB	Ordenação B, A, AB
MINITAB	A	$H_0^{(1)}$, $H_0^{(3)}$	$H_0^{(2)}$, $H_0^{(3)}$
NTIA	B	$H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H_0^{(6)}$
S-PLUS	AB	$H_0^{(7)}$	$H_0^{(7)}$
STATISTICA	A	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$
	B	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H^{(5)}$, $H_0^{(6)}$
BMDP	AB	$H_0^{(7)}$	$H_0^{(7)}$
SPSS	A	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(2)}$, $H_0^{(3)}$
	B	$H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$
	AB	$H_0^{(7)}$	$H_0^{(7)}$

Quadro VI: Análise de variância do modelo (2.1) fornecida por diversos softwares, a partir dos dados do Quadro II.

Software	Variações	Hipóteses Testadas	
	Consideradas	Ordenação A, B, AB	Ordenação B, A, AB
MINITAB	A	$H_0^{(1)}$	$H_0^{(2)}$
	B	$H_0^{(5)}$	$H_0^{(4)}$
NTIA	AB	$H_0^{(7)}$	$H_0^{(7)}$
BMDP	A	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$
	B	$H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H^{(5)}$, $H_0^{(6)}$
	AB	$H_0^{(7)}$	$H_0^{(7)}$
SPSS	A	$H_0^{(1)}$, $H_0^{(2)}$, $H_0^{(3)}$	$H_0^{(2)}$, $H_0^{(3)}$
	B	$H_0^{(5)}$, $H_0^{(6)}$, $H_0^{(8)}$	$H_0^{(4)}, H_0^{(5)}, H_0^{(6)}, H^{(8)}$
	AB	$H_0^{(7)}$	$H_0^{(7)}$
S-PLUS	A	$H_0^{(1)}$, $H_0^{(3)}$	$H_0^{(2)}$, $H_0^{(3)}$
	B	$H_0^{(5)}$, $H_0^{(6)}$	$H_0^{(4)}$, $H_0^{(6)}$
	AB	$H_0^{(7)}$	$H_0^{(7)}$
STATISTICA	não disponível		

4. Conclusões

Face aos resultados obtidos, concluiu-se que:

• dos *softwares* estudados, apenas o SPSS teve os mesmos resultados do *software* SAS na obtenção das somas de quadrados para dados desbalanceados com e sem caselas vazias;

• o MINITAB e o NTIA fornecem em suas saídas, somas de quadrados dos tipos seqüenciais e ajustadas, equivalentes às somas de quadrados dos Tipos I e III fornecidas pelo procedimento GLM do SAS. Entretanto, se há caselas vazias, o MINITAB e o NTIA só forneceram as somas de quadrados do tipo seqüenciais, equivalentes às somas de quadrados do Tipo I;

• o *software* S-PLUS, fornece em sua saída somas de quadrados dos tipos seqüenciais e ajustadas tanto para o caso em que todas as caselas são ocupadas, quanto para o caso em que há caselas vazias. Fornecendo assim, somas de quadrados equivalentes às somas de quadrados dos Tipos I e III do SAS;

• o STATISTICA, quando as caselas estão ocupadas, fornece somas de quadrados para os efeitos principais e interações, A, B e AB. As suas somas de quadrados são relacionadas às hipóteses sobre médias ponderadas não ajustadas, médias ponderadas ajustadas e médias não ponderadas. Assim, a ordem de entrada dos fatores no modelo não influi na obtenção das somas de quadrados. Agora, se existem caselas vazias, então, o STATISTICA emite a mensagem "DESIGN INCOMPLETE; REGRESSION APPROACH NOT AVAILABLE" e não fornece essas somas de quadrados;

• da mesma forma que o *software* STATISTICA, o BMDP não depende da ordenação do modelo para fornecer suas somas de quadrados. Assim, tanto para o caso em que todas as caselas são ocupadas, quanto para o 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.

Com base nas conclusões obtidas, 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 que esses *softwares* estão testando pode induzir à 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.

Abstract. The interpretation problem of the tested hypothesis in the models with two crossed factors and unbalanced data was widely argued on the basis of GLM procedure of the SAS (Iemma [1] and Mondardo [2]). In this context, the aim of this work is to compare the results of others softwares with those results provided by GLM procedure of the SAS.

Referências

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

[2] M. Mondardo, "Estimabilidade de Funções Paramétricas com Dados Desbalanceados Através do PROC-GLM do SAS: Aplicações à Pesquisa Agropecuária", Dissertação de Mestrado, ESALQ, USP, Piracicaba, SP, 1994.

[3] S.M. Oikawa, "Hipóteses Estatísticas com Dados Desbalanceados nos Modelos de Efeitos Fixos Hierarquizados em Presença ou não de Esquema Fatorial", Tese de Doutorado, ESALQ, USP, Piracicaba, SP, 1998.

[4] S.R. Searle, "Linear Models", John Wiley & Sons, New York, 1971.

[5] X, "SAS User's guide: Statistics version 6 edition ", Cary, SAS Institute, 1990, 846p.