**HIPÓTESES ESTATÍSTICAS COM DADOS DESBALANCEADOS
NOS MODELOS DE EFEITOS FIXOS HIERARQUIZADOS
EM PRESENÇA OU NÃO DE ESQUEMA FATORIAL**

SÉRGIO MINORU OIKAWA

Estatístico

Orientador: Prof. Dr. Antonio Francisco Iemma

Tese apresentada à Escola Superior de
Agricultura “Luiz de Queiroz”, Univer-
sidade de São Paulo, para obtenção do
título de Doutor em Agronomia. Área
de concentração: Estatística e Experi-
mentação Agrônômica.

PIRACICABA

Estado de São Paulo - Brasil

Agosto - 1998

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - Campus "Luiz de Queiroz"/USP

Oikawa, Sérgio Minoru

Hipóteses estatísticas com dados desbalanceados nos modelos de efeitos fixos hierarquizados em presença ou não de esquema fatorial / Sérgio Minoru Oikawa. - - Piracicaba, 1998.

145 p.

Tese (doutorado) - - Escola Superior de Agricultura Luiz de Queiroz, 1998.
Bibliografia.

1. Análise de variância 2. Estatística computacional 3. Hipótese estatística 4. Modelo Gauss-Markov 5. Modelo matemático 6. Programa de computador I. Título

CDD 519.535

À DEUS,
pela força divina nos
momentos mais difíceis,
DEDICO

À minha esposa,
Luciana,
e à minha filha,
Tatiana,
OFEREÇO

AGRADECIMENTOS

- Ao Dr. *Antonio Francisco Lemma*, Professor Titular do Departamento de Matemática e Estatística da ESALQ-USP, pela constante e eficiente orientação.
- Aos *Professores* do Departamento de Matemática e Estatística da ESALQ-USP, pelos ensinamentos, apoios e incentivos oferecidos.
- À *CAPES/PICD*, pelo auxílio financeiro concedido para a realização do programa de Doutorado.
- À *CPPSE/EMBRAPA*, São Carlos - SP, por ter concedido os dados.
- Aos *Colegas* do Curso de Pós-Graduação em Estatística e Experimentação Agronômica pela amizade.
- Aos *Colegas* do Departamento de Matemática da FCT-UNESP, pelo apoio, incentivo e compreensão.
- Aos *Funcionários* do Departamento de Matemática e Estatística da ESALQ-USP, pelas gentilezas e agradável convívio.
- À *todos* aqueles que, de uma forma ou outra, contribuíram para a realização deste trabalho.

SUMÁRIO

1	INTRODUÇÃO.....	01
2	REVISÃO DE LITERATURA.....	06
3	METODOLOGIA.....	37
3.1	Modelo Linear de Gauss-Markov Normal.....	37
3.2	Modelos com Estrutura do Tipo Estritamente Hierárquica.....	38
3.2.1	Modelo de Médias de Caselas (Modelo-M).....	40
3.2.2	Modelo Superparametrizado (Modelo-S).....	42
3.2.3	Obtenção das Somas de Quadrados.....	44
3.3	Modelos com Estrutura do Tipo Fatorial-Hierárquica.....	49
3.3.1	Modelo de Médias de Caselas (Modelo-M).....	50
3.3.2	Modelo Superparametrizado (Modelo-S).....	51
3.3.3	Obtenção das Somas de Quadrados.....	52
3.4	Hipóteses Estatísticas Usualmente Testadas Pelo SAS-GLM.....	57
3.4.1	Hipóteses Testadas nos Modelos com Três Fatores Hierarquizados.....	57
3.4.1.1	Hipóteses sobre o fator A.....	58
3.4.1.2	Hipóteses sobre o fator B aninhado sob o fator A.....	59
3.4.1.3	Hipótese sobre o fator C aninhado sob o fator B(A).....	60
3.4.2	Hipóteses Testadas nos Modelos com Estrutura do Tipo Fatorial-Hierárquica.....	61
3.4.2.1	Hipóteses sobre o fator A.....	61

3.4.2.2	Hipóteses sobre o fator B.....	65
3.4.2.3	Hipóteses sobre a Interação A*B.....	66
3.4.2.4	Hipóteses sobre o fator C aninhado sob o fator B.....	68
3.4.2.5	Hipótese sobre a interação A*C(B).....	70
4	RESULTADOS E DISCUSSÃO.....	72
4.1	Modelos com Três Fatores Hierarquizados.....	72
4.1.1	Obtenção das Somas de Quadrados.....	73
4.1.2	Hipóteses Testadas Pelo SAS-GLM nos Modelos com Três Fatores Hierarquizados.....	75
4.1.2.1	Hipóteses sobre o fator A.....	75
4.1.2.2	Hipóteses sobre o fator B aninhado sob o fator A.....	81
4.1.2.3	Hipóteses sobre o fator C aninhado sob o fator B(A).....	85
4.1.3	Hipóteses Testadas Por Outros <i>Softwares</i> nos Modelos com Três Fatores Hierarquizados.....	87
4.1.3.1	MINITAB.....	87
4.1.3.2	NTIA.....	88
4.1.3.3	STATGRAPHICS.....	90
4.1.3.4	SAEG.....	90
4.1.3.5	SANEST.....	91
4.1.3.6	GLIM.....	91
4.1.3.7	STATISTICA.....	92
4.1.3.8	BMDP.....	93
4.2	Modelos com Estrutura do Tipo Fatorial-Hierárquica.....	95

4.2.1	Obtenção das Somas de Quadrados.....	96
4.2.2	Hipóteses Testadas Pelo SAS-GLM nos Modelos com Estrutura do Tipo Fatorial-Hierárquica.....	99
4.2.2.1	Hipóteses sobre o fator A.....	99
4.2.2.2	Hipóteses sobre o fator B.....	105
4.2.2.3	Hipóteses sobre a interação A*B.....	109
4.2.2.4	Hipóteses sobre o fator C(B).....	111
4.2.2.5	Hipótese sobre a interação A*C(B).....	116
4.2.3	Hipóteses Testadas Por Outros <i>Softwares</i> nos Modelos com Estrutura do Tipo Fatorial-Hierárquica.....	123
4.2.3.1	MINITAB.....	123
4.2.3.2	NTIA.....	124
4.2.3.3	STATGRAPHICS.....	126
4.2.3.4	SAEG.....	126
4.2.3.5	SANEST.....	126
4.2.3.6	GLIM.....	126
4.2.3.7	STATISTICA.....	128
4.2.3.8	BMDP.....	129
5	CONCLUSÕES.....	135
	REFERÊNCIAS BIBLIOGRÁFICAS.....	140

HIPÓTESES ESTATÍSTICAS COM DADOS DESBALANCEADOS NOS MODELOS DE EFEITOS FIXOS HIERARQUIZADOS EM PRESENÇA OU NÃO DE ESQUEMA FATORIAL

Autor: Sérgio Minoru Oikawa

Orientador: Prof. Dr. Antonio Francisco Iemma

RESUMO

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 sua grande rapidez e seu baixo custo operacional por unidade aritmética. Tais características, bem como a facilidade de acesso, fizeram com que o número de usuários de *softwares* estatísticos crescesse consideravelmente.

Infelizmente, as dificuldades encontradas pelos usuários de *softwares* estatísticos têm sido agravadas pela escassez de informações detalhadas sobre as interpretações de hipóteses estatísticas em modelos mais complexos com dados desbalanceados, tanto na literatura como nos manuais de utilização. Assim, a ausência de manuais bem documentados na maioria dos *softwares*, tem induzido muitos usuários a cometerem erros de interpretação das hipóteses que podem comprometer os resultados de sua pesquisa.

Nesse contexto, um dos objetivos naturais desse trabalho, foi o estudo dos modelos de efeitos fixos hierarquizados em presença ou não de esquema fatorial com

dados desbalanceados. Para esses modelos, formalizaram-se os termos das respectivas somas de quadrados e hipóteses estatísticas mais usuais que são testadas para os efeitos principais e interações. Discutiu-se, também, o problema da interpretação de hipóteses associadas às somas de quadrados fornecidas por alguns dos sistemas computacionais estatísticos universalmente consagrados.

Visando ampliar os horizontes dos profissionais das ciências aplicadas, usuários de sistemas computacionais estatísticos disponíveis em seus centros de pesquisa, compararam-se as performances de alguns sistemas como o SAS, o MINITAB, o GLIM, o BMDP, o STATGRAPHICS, o STATISTICA, o SAEG, o SANEST e o NTIA.

Face aos resultados, concluiu-se que:

- Quando os dados são desbalanceados com todas as caselas ocupadas, porém os níveis do fator C são diferentes, as somas de quadrados referentes ao fator A fornecidas pelo SAS-GLM, não testam as hipóteses sobre as médias não ponderadas proposto por Yates (1934), mas uma hipótese não usual, gerada a partir de funções estimáveis complexas do tipo III.

- As somas de quadrados parciais referentes ao fator A fornecidas pelo NTIA, também não testam as hipóteses sobre as médias não ponderadas, quando os dados são desbalanceados e os níveis do fator C são diferentes.

- A ocorrência de dados desbalanceados em presença de caselas vazias pode trazer sérios transtornos aos pesquisadores das ciências aplicadas, com 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.

**STATISTICAL HYPOTHESES WITH UNBALANCED
DATA IN THE MODELS OF NESTED FIXED EFFECTS
IN PRESENCE OR NOT OF SCHEME FACTORIAL**

Author: Sérgio Minoru Oikawa

Adviser: Prof. Dr. Antonio Francisco Iemma

SUMMARY

The statistical softwares became an important and indispensable tool in the statistical analysis of data, mainly, due to the capacity of the computers today, such as its great speed and its low operational cost for arithmetic unit. Such characteristic, as well as the access easiness, caused the number of users of statistical softwares to grow considerably.

Unfortunately, the difficulties found by the users of statistical softwares have been worsened by the lack of detailed information about the interpretations of statistical hypotheses in more complex models with unbalanced data, either in the literature or in the use manuals. Thus, the absence of manuals well documented in most of the softwares, has been inducing many users they to make misinterpretations of the hypotheses which can commit the results of their research.

In that context, one of the natural objectives of this work, was the study of the models of nested fixed effects in presence or not of scheme factorial with unbalanced data. For those models, the terms of the respective sums of squares and more usual sta-

tistical hypotheses than they are tested for the main effects and interactions were formalized. The problem of the interpretation of hypotheses associated to the sums of squares provided universally by some of the systems statistical computacionais consecrated were also discussed.

Seeking to enlarge the professionals' of the applied sciences horizons, users of available statistical computacionais systems in their research centers, the performance of some systems as the SAS ones, MINITAB, GLIM, BMDP, STATGRAPHICS, STATISTICA, SAEG, SANEST and NTIA were compared.

According to the results, it was found that:

- When the data are unbalanced with all the cells filled, but the levels of the factor C are different, then the sums of squares referring to the factor A supplied by the SAS-GLM don't test the hypotheses on the non pondered averages proposed by Yates (1934), but a unusual hypothesis, generated from complex estimable functions of the type III.

The sums of squares partial referring the factor A supplied by NTIA, also don't test the hypotheses on the non pondered averages, when the data are unbalanced and the levels of the factor C are different.

- The occurrence of unbalanced data in presence of empty cells can bring serious upset to the researchers of the applied sciences, regarding the interpretations of statistical hypotheses, because in most of the cases, the lack of an explicit documentation about what those softwares are calculating, can induce the taking of incorrect decisions.

1 INTRODUÇÃO

Atualmente, os *softwares* estatísticos tornaram-se uma ferramenta importante e indispensável na análise estatística de dados, especialmente nas ciências aplicadas. Segundo Searle (1987), os *softwares* estatísticos, hoje disponíveis, são capazes de realizar cálculos aritméticos complexos que eram totalmente inconcebíveis há trinta anos atrás. A capacidade dos computadores de hoje, sua grande rapidez e seu baixo custo operacional por unidade aritmética, eram características totalmente inimagináveis para muitos estatísticos da década de 50. Tais características marcantes, bem como a facilidade de acesso, fizeram com que o número de usuários de *softwares* estatísticos crescesse consideravelmente.

Esse fato acabou gerando um sério problema, motivado pela utilização muitas vezes inadequada dos *softwares* estatísticos. Isto porque os resultados de uma análise de dados balanceados ($n_{ijk} = k$), geralmente, são resumidos em um único quadro de análise da variância, onde cada soma de quadrados obtida de partições ortogonais está associada a uma particular hipótese estatística. Essas hipóteses são bem definidas e geralmente não causam confusão, entre os usuários, sobre o que está sendo testado. Já os resultados de uma análise de dados desbalanceados em presença ou não de caselas vazias ($n_{ijk} \neq k$) não podem, em geral, ser resumidos num único quadro de análise da variância,

pois existem diferentes formas de partições das somas de quadrados e, conseqüentemente, isso pode causar confusão quanto à forma das hipóteses testadas.

Nesse contexto, segundo Iemma (1993), é importante enfatizar sobre as interpretações das verdadeiras hipóteses estatísticas que são testadas através das somas de quadrados, obtidas por diversos métodos disponíveis na literatura. Assim, 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, podem testar quatro diferentes hipóteses.

Tomando-se como exemplo um modelo com dois fatores cruzados e interação, se os dados são balanceados, não há dificuldades para as interpretações de tais hipóteses, pois elas são todas equivalentes. Nesse caso, os usuários, especialmente os pesquisadores das ciências aplicadas, podem interpretar facilmente as hipóteses que são testadas por diversos *softwares* estatísticos. Assim, as análises estatísticas podem ser realizadas através do *software* de sua preferência e/ou disponibilidade.

Se, entretanto, os dados são desbalanceados em presença ou não de caselas vazias, os métodos disponíveis fornecem diferentes somas de quadrados e, portanto, testam diferentes hipóteses. Desse modo, os pesquisadores não iniciados na análise de dados desbalanceados, podem estar testando hipóteses completamente diferentes daquelas que julgam testar, alterando sensivelmente as interpretações dos resultados de seus experimentos. Existem na literatura inúmeros artigos que tratam sobre o assunto, porém, em casos de dados desbalanceados em presença ou não de caselas vazias, ainda há muitas

controvérsias sobre a utilização de uma metodologia unificada. Evidentemente, isso tem refletido sobre os *softwares* estatísticos que utilizam diferentes métodos e, portanto, fornecem diferentes resultados para o mesmo conjunto de dados.

A situação pode agravar-se, ainda, quando são utilizados modelos mais parametrizados envolvendo vários fatores cruzados e hierárquicos com interações. Dallal (1992) apresenta tal situação usando como exemplo um modelo com três fatores (A, B, C) e interação $A*B$, onde os fatores A e B são cruzados e ele introduz um terceiro fator C hierarquizado (aninhado) sob a interação $A*B$. Para esse modelo constata que, mesmo os dados apresentando desbalanceamento somente nos níveis do fator $C(A*B)$, tanto o procedimento GLM do SAS, versão 6.04, como o procedimento MANOVA (*Multiple Analysis of Variance*) do SPSS (*Statistical Package for the Social Sciences*), versão 4.0, além de fornecerem resultados diferentes para as somas de quadrados, não fornecem qualquer informação nos manuais de utilização a respeito das hipóteses testadas.

Agora, se os dados são desbalanceados em presença de caselas vazias, então, as análises podem tornar-se ainda mais complexas, pois as somas de quadrados fornecidas pelos *softwares* estatísticos, em particular o SAS-GLM, testam hipóteses, algumas vezes, geradas a partir de funções estimáveis complexas de difícil interpretação.

Infelizmente, as dificuldades encontradas pelos usuários de *softwares* estatísticos têm sido agravadas pela escassez de informações detalhadas sobre as interpretações das hipóteses estatísticas nos modelos mais complexos, tanto na literatura como nos manuais de utilização, especialmente em relação aos modelos com estruturas hierarquizadas em presença ou não de esquemas fatoriais. Assim, a ausência de manuais bem docu-

mentados na maioria dos *softwares*, tem induzido muitos usuários a cometerem erros de interpretação que podem comprometer os resultados de sua pesquisa.

Nesse contexto, um dos objetivos naturais desse trabalho, é o estudo dos modelos hierarquizados em presença ou não de esquema fatorial, com dados desbalanceados. Discute-se o problema da interpretação das hipóteses estatísticas associadas às somas de quadrados fornecidas por alguns dos sistemas computacionais estatísticos universalmente consagrados. Com essa conduta, espera-se amenizar os problemas levantados por Dallal (1992) e Searle (1994).

Visando ampliar os horizontes dos profissionais das ciências aplicadas, usuários de sistemas computacionais estatísticos disponíveis em seus centros de pesquisa, comparam-se, neste estudo, sem o apelo de competição, as performances de alguns sistemas como o SAS (*Statistical Analysis System*) - versão 6.11; o MINITAB (*Minitab Statistical Softwares*) - versão 11.1; o GLIM (*Generalised Linear Interactive Modeling*) - versão 4.0; o STATGRAPHICS (*Statistical Graphics System*) - versão 7.0; o BMDP (*Biomedical Computer Programs*) - versão PC90; o STATISTICA (*Integrated System of statistical Analysis, Generation of Graphs and Management of Bases of Data*) - versão 5.0; o SAEG (Sistema para Análises Estatísticas e Genéticas) - versão 5.0; o SANEST (Sistema de Análise Estatística) - versão 2.1 e o NTIA (Núcleo Tecnológico de Informática para a Agricultura) - versão 4.2.2, em relação ao tema deste estudo.

Um outro fato a ser ressaltado é o de que nesse estudo os *softwares* são abordados do ponto de vista do usuário e não do especialista. Nesse contexto, eles são apenas utilizados para analisar os exemplos numéricos apresentados no item Resultados e

Discussão, e não constam no item Metodologia pois não há aqui o objetivo de descrevê-los e analisá-los com detalhes.

2 REVISÃO DE LITERATURA

Segundo Herr (1986), a história da análise de dados desbalanceados pode ser descrita desde as publicações originais de Yates (1933 e 1934), Hendricks (1934), Snedecor (1934), Snedecor & Cox (1935), Comstock (1943), Hazel (1946), Patterson (1946), Vajda (1947), Yates (1947), Finney (1948), Henderson (1953), Kramer (1955) até o início dos anos sessenta com a revolução computacional. Na década de 60, surgem entre outros, os artigos de Elston & Bush (1964), Searle (1965), Gosslee & Lucas (1965), Federer & Zelen (1966) e Overall & Spiegel (1969).

Dentre os artigos citados por Herr (1986), merecem destaque os artigos de Yates (1933 e 1934), Henderson (1953) e Overall & Spiegel (1969), cujos métodos estão incorporados pelo procedimento GLM do SAS.

Sem dúvida, os artigos propostos por Yates (1933 e 1934) são um marco no estudo da estimação e dos testes de hipóteses com dados desbalanceados. Por exemplo, Yates (1934) propõe soluções razoáveis para o problema de classificações duplas cruzadas (Two-Way),

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

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

De acordo com Herr (1986), praticamente todos os métodos hoje utilizados são derivados do artigo de Yates (1934). Ali, o autor propôs e discutiu os seguintes métodos:

- Método para classes proporcionais (Y_1).

Incorporado pelo SAS-GLM, fornece somas de quadrados do tipo I: $R(\alpha | \mu)$ ou $R(\beta | \mu)$, e testa hipóteses do tipo I sobre as médias ponderadas não ajustadas para outros fatores e interações (Iemma, 1993).

- Método do ajustamento de constantes (Y_2).

Incorporado pelo SAS-GLM, fornece somas de quadrados do tipo II: $R(\alpha | \mu, \beta)$ e $R(\beta | \mu, \alpha)$, e testa hipóteses do tipo II sobre as médias ponderadas ajustadas para todos os fatores e interações, exceto interações e/ou fatores hierarquizados que envolvam o fator de interesse.

- Método dos quadrados de médias ponderadas (Y_3).

Também incorporado pelo SAS-GLM, corresponde às somas de quadrados do tipo III, $R[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*]$ e $R[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*]$. Segundo Iemma (1993), testa hipóteses do tipo III sobre as médias não ponderadas, se os dados são desbalanceados com todas as caselas ocupadas.

Conforme Herr (1986), Yates (1934) declara explicitamente o que ele quer dizer com relação à avaliação dos efeitos principais. Está claro em sua discussão que em

testando, por exemplo, a hipótese H_0 : não há diferenças entre efeitos principais de linhas, ele está testando que as médias não ponderadas de linhas das médias de caselas, tomadas sobre todas as colunas são iguais. Sem dúvida, essa é a hipótese testada pelo método dos quadrados de médias ponderadas. Contudo, teria sido mais útil se tivesse declarado a hipótese através de fórmulas explícitas em lugar de simples descrição verbal. Yates (1934) também distingue claramente entre o modelo aditivo (sem interação) e o modelo completo (superparametrizado).

Segundo Yates (1934), os métodos de estimação e testes de hipóteses em modelos com dois fatores cruzados, diferem de acordo com a suposição de existência ou não da interação. Assim, se for considerado o modelo aditivo, então o método do ajustamento de constantes fornece estimativas e testes de hipóteses “eficientes” e mais poderosos que o método dos quadrados de médias ponderadas. Caso contrário, o método dos quadrados de médias ponderadas é preferível.

No caso de números de classes proporcionais, por exemplo, as médias dos totais marginais do fator A são estimativas eficientes dos efeitos principais de A, avaliados sobre qualquer número de classes em B. Nesse caso, os totais marginais de A não contêm o mesmo número de observações, porém as somas de quadrados apropriadas para testar os efeitos do fator A podem ser calculadas pelo método descrito para classificações simples (One-Way) com dados desbalanceados. Os efeitos do fator B podem ser estimados e testados de modo semelhante e, finalmente, a interação pode ser testada subtraindo-se as duas somas de quadrados (A e B) da soma de quadrados dos totais entre classes, já que as somas de quadrados para efeitos de A, efeitos de B e interação são

aditivos. De acordo com Herr (1986), Yates (1934) está referindo-se as médias dos totais marginais do fator A como médias ponderadas. Assim, quando ele afirma que as três somas de quadrados são aditivas, ele está definindo os efeitos principais através de médias ponderadas das médias de caselas.

Segundo Iemma (1993, 1995-a e b), outro artigo antológico foi publicado por Henderson (1953), que propõe três métodos para a obtenção das somas de quadrados utilizadas no estudo de componentes de variância em modelos mistos e aleatórios. Como as somas de quadrados obtidas pelos métodos 1 e 3 são usadas para testes de hipóteses em modelos de efeitos fixos, Speed *et al.* (1978) incorporam tais métodos ao rol dos mais utilizados na análise de variância com dados desbalanceados. Para aplicar os métodos propostos, Henderson (1953), utiliza como exemplo um modelo com três fatores cruzados (A, H, S) e interação H*S,

$$y_{ijkt} = \mu + a_i + h_j + s_k + (hs)_{jk} + e_{ijkt} ,$$

$$i = 1, 2, \dots, p ; j = 1, 2, \dots, q ; k = 1, 2, \dots, r ; t = 1, 2, \dots, n_{ijk} .$$

Os métodos de Henderson (1953), utilizados na análise de modelos de efeitos fixos, são:

- Método 1 de Henderson (H_1).

No método H_1 , as somas de quadrados dos efeitos de A, R ($a \mid \mu$), efeitos de H, R ($h \mid \mu$), efeitos de S, R ($s \mid \mu$) e da interação H*S, R [$\mu, h, s, (hs)$] - R (μ, h) -

- $R(\mu, s) + R(\mu)$, derivam do método Y_1 de Yates e, portanto, as somas de quadrados dos efeitos principais testam hipóteses do tipo I sobre as médias ponderadas não ajustadas. Entretanto, segundo Searle (1971), quando os dados são desbalanceados, a quantidade associada à interação $H \times S$, obtida por subtração, não corresponde à hipótese usual sobre a interação.

- Método 3 de Henderson (H_3).

O método H_3 pode ser usado com qualquer tabela de análise de variância (método Y_2 , método Y_3 etc.). No exemplo de Henderson (1953), ele obteve as somas de quadrados $R[a | \mu, h, s, (hs)]$, $R(h | \mu, a, s)$, $R(s | \mu, a, h)$ e $R[(hs) | \mu, a, h, s]$ que correspondem ao método Y_2 de Yates. Assim, as somas de quadrados dos efeitos principais testam hipóteses do tipo II sobre as médias ponderadas ajustadas para todos os fatores e interação, exceto interação que envolve o fator de interesse. Segundo Speed *et al.* (1978), o método H_3 é geralmente associado às somas de quadrados dos efeitos principais oriundas de ordenações *a priori* de interesse, por exemplo, $R(a | \mu)$, $R(h | \mu, a)$, $R(s | \mu, a, h)$ e $R[(hs) | \mu, a, h, s]$, portanto, corresponde a um caso particular do método Y_2 de Yates. Sendo assim, testa hipóteses sobre as médias ponderadas não ajustadas e hipóteses sobre as médias ponderadas ajustadas para todos os fatores que precedem o fator de interesse.

Outro trabalho de destaque na literatura é o artigo de Overall & Spiegel (1969), no qual os autores propõem três métodos para a análise de dados desbalanceados. Nesse artigo eles generalizam os principais métodos apresentados nos artigos ante-

riores, utilizando modelos reparametrizados de posto completo. Como aplicação, consideram exemplos com dois modelos: um com dois fatores cruzados (A, B) e interação A*B e o outro com três fatores cruzados (A, B, C) e interações duplas (A*B, A*C, B*C),

$$y_{ijk} = \mu^* + \alpha_i^* + \beta_j^* + (\alpha\beta)_{ij}^* + \varepsilon_{ijk} ,$$

$$i = 1, 2, \dots, (a - 1) ; j = 1, 2, \dots, (b - 1) ; k = 1, 2, \dots, n_{ij} \quad e$$

$$y_{ijkt} = \mu^* + \alpha_i^* + \beta_j^* + \gamma_k^* + (\alpha\beta)_{ij}^* + (\alpha\gamma)_{ik}^* + (\beta\gamma)_{jk}^* + \varepsilon_{ijkt} ,$$

$$i = 1, \dots, (a - 1) ; j = 1, \dots, (b - 1) ; k = 1, \dots, (c - 1) ; t = 1, \dots, n_{ijk} .$$

Segundo Iemma (1993, 1995-a e b), em 1976 o SAS-GLM incorpora esses três métodos, que estão consagrados na literatura como sendo propostos por Overall & Spiegel (1969), embora Yates (1934) e Henderson (1953) os tenham propostos anteriormente. Basicamente, tais métodos podem ser assim resumidos:

- Método dos mínimos quadrados completos (OS₁).

Esse método é precisamente o método Y₃ de Yates e, portanto, as somas de quadrados dos efeitos principais testam hipóteses do tipo III, sobre as médias não ponderadas. Por exemplo, no modelo com dois fatores cruzados e interação, se os dados são desbalanceados com todas as caselas ocupadas, o método OS₁ fornece somas de quadrados, $R[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*]$, $R[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*]$ e $R[(\alpha\beta)^* | \mu^*, \alpha^*, \beta^*]$.

- Método do planejamento experimental (OS₂).

Corresponde ao método Y₂ de Yates, onde as somas de quadrados dos efeitos principais são ajustadas para todos os fatores e interações, exceto interações que envolvem o fator de interesse, $R[\alpha^* | \mu^*, \beta^*]$, $R[\beta^* | \mu^*, \alpha^*]$ e $R[(\alpha\beta)^* | \mu^*, \alpha^*, \beta^*]$. Assim, testa hipóteses do tipo II sobre as médias ponderadas ajustadas.

- Método da ordenação *a priori* (OS₃).

Corresponde a um caso particular do método Y₂ de Yates e utiliza, *a priori*, ordenações paramétricas de interesse. Assim, se os parâmetros entram no modelo na ordem (A, B, AB), as somas de quadrados obtidas são: $R[\alpha^* | \mu^*]$, $R[\beta^* | \mu^*, \alpha^*]$ e $R[(\alpha\beta)^* | \mu^*, \alpha^*, \beta^*]$. De modo análogo, para a ordem (B, A, AB), tem-se: $R[\beta^* | \mu^*]$, $R[\alpha^* | \mu^*, \beta^*]$ e $R[(\alpha\beta)^* | \mu^*, \alpha^*, \beta^*]$.

Na década de 70, em virtude das confusões na interpretação de hipóteses estatísticas nos experimentos com dados desbalanceados, bem como à disponibilidade de inúmeros *softwares* estatísticos que não fornecem resultados equivalentes 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 realizadas através de diferentes *softwares* estatísticos, os resultados obtidos para as somas de quadrados, além de não serem os mesmos, algumas vezes, são incorretos. Para ilustrar o problema, utiliza como exemplo um modelo com dois fatores cruzados (A, B) e interação,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2, \dots, s ; j = 1, 2, \dots, r ; k = 1, 2, \dots, n_{ij} .$$

Nesse estudo, o autor verifica que o procedimento ANOVA (*Analysis of Variance*) do SAS (versão 1971) fornece soma de quadrados negativa para a interação. Segundo ele, não há nenhuma informação na documentação do SAS-ANOVA a respeito da causa do problema. Francis (1973) considera, também, os *softwares*: BMDX64 (*Biomedical Computer Programs* - versão 1970), o procedimento MANOVA do OSIRIS-II (*An Integrated Collection of Computer Programs for the Management and Analysis of Social Science Data* - versão 1970) e o “CAROLINA-MANOVA”, onde segundo Cramer (1976), autor do *software* CAROLINA-MANOVA, o termo correto é MANOVA (*Multivariate Analysis of Variance Programs* - versão 1967). Constata que os resultados das somas de quadrados dos efeitos principais não são equivalentes para os três *softwares*. Segundo o autor, apenas a documentação do BMDX64 esclarece como as somas de quadrados são calculadas, possibilitando que qualquer usuário interprete os resultados e faça um melhor uso do programa. As documentações do “CAROLINA” e OSIRIS-II foram de pouca valia na tentativa de interpretar os resultados obtidos, pois ambos explicam como escolher entre as várias opções disponíveis, mas não explicam como os cálculos são realizados. Finalmente, conclui que a soma de quadrados negativa, fornecida pelo SAS-ANOVA é devida a não-ortogonalidade entre os efeitos principais, portanto, a soma de quadrados da interação não deve ser obtida por subtração.

É interessante ressaltar que as somas de quadrados dos efeitos principais fornecidas pelo SAS-ANOVA são calculadas através do método Y_1 de Yates. As somas de quadrados fornecidas pelos *softwares* “CAROLINA” e OSIRIS-II dependem das ordenações *a priori* de interesse. Por exemplo, o “CAROLINA” considera a ordenação (A, B, A*B) e fornece somas de quadrados, $R(\alpha | \mu)$ e $R(\beta | \mu, \alpha)$. O OSIRIS-II considera a ordenação (A*B, B, A), fornecendo somas de quadrados, $R(\beta | \mu)$ e $R(\alpha | \mu, \beta)$ e, portanto, eles utilizam o método OS_3 de Overall & Spiegel (1969). Já as somas de quadrados fornecida pelo BMDX64 utilizam o método OS_1 , fornecendo somas de quadrados, $R[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*]$ e $R[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*]$.

Segundo Carlson & Timm (1974), os três métodos sugeridos por Overall & Spiegel (1969), baseados em modelos reparametrizados de posto completo, podem estar confundindo os pesquisadores não familiarizados com a teoria clássica de modelos lineares. Isto porque os modelos reparametrizados de posto completo não especificam claramente a natureza de diferentes hipóteses testadas por diversos métodos. Assim, Carlson & Timm (1974) especificam explicitamente a natureza das hipóteses testadas pelos três métodos de Overall & Spiegel (1969), através dos modelos de médias de caselas e associam tais hipóteses às somas de quadrados do modelo reparametrizado de posto completo. Segundo os autores, Searle (1971) já havia discutido as hipóteses testadas em modelos de posto incompleto, utilizando o conceito da redução nas somas de quadrados através da notação-R ($\cdot$), fornecendo alguma visão das diferenças entre diferentes métodos de análise. No entanto, os autores afirmam que os modelos de médias de caselas esclarecem melhor a natureza de diferentes hipóteses que são testadas por diver-

soos métodos. Consideram um exemplo com dois fatores cruzados (A, B) e interação, onde o fator A tem dois níveis e o fator B tem três níveis, descrito através do modelo de médias de caselas:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2 ; j = 1, 2, 3 ; k = 1, 2, \dots, n_{ij} .$$

Assim, se os pesquisadores têm interesse em testar hipóteses do tipo III sobre as médias não ponderadas, então, as hipóteses dos efeitos principais correspondentes ao método OS₁ de Overall & Spiegel (1969) são dadas por:

$$H_A : \bar{\mu}_{1\bullet} = \bar{\mu}_{2\bullet} \Leftrightarrow \left(\sum_{j=1}^3 \mu_{1j} \right) / 3 = \left(\sum_{j=1}^3 \mu_{2j} \right) / 3$$

$$H_B : \bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3} \Leftrightarrow \left(\sum_{i=1}^2 \mu_{i1} \right) / 2 = \left(\sum_{i=1}^2 \mu_{i2} \right) / 2 = \left(\sum_{i=1}^2 \mu_{i3} \right) / 2 ,$$

ou equivalentemente,

$$H_A : (\mu_{11} + \mu_{12} + \mu_{13}) / 3 = (\mu_{21} + \mu_{22} + \mu_{23}) / 3 ,$$

$$H_B : (\mu_{11} + \mu_{21}) / 2 = (\mu_{12} + \mu_{22}) / 2 = (\mu_{13} + \mu_{23}) / 2 .$$

Segundo os autores, as hipóteses H_A e H_B podem ser representadas na forma geral,

$$H_0 : B'\mu = h,$$

onde B' é uma matriz de dimensão $(s \times ab)$ de constantes conhecidas, de posto linha completo, h é um vetor de dimensão $(s \times 1)$ de constantes conhecidas (geralmente igual a zero), s corresponde aos graus de liberdade (g.l.) associados aos efeitos principais e ab corresponde ao número total de parâmetros do modelo de médias de caselas. Nesse caso, considerando-se o fator A com $s = (a - 1) = 1$ g.l. e o fator B com $s = (b - 1) = 2$ g.l., têm-se:

$$H_A : \begin{bmatrix} 1/3 & 1/3 & 1/3 & -1/3 & -1/3 & -1/3 \end{bmatrix} \mu = 0,$$

$$H_B : \begin{bmatrix} 1/2 & 0 & -1/2 & 1/2 & 0 & -1/2 \\ 0 & 1/2 & -1/2 & 0 & 1/2 & -1/2 \end{bmatrix} \mu = 0,$$

onde,

$$\mu' = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{21} & \mu_{22} & \mu_{23} \end{bmatrix}.$$

Segundo Iemma (1987), dado o modelo $y = W\mu + \varepsilon$; $\varepsilon \sim N(\phi, I\sigma^2)$, se $B'\mu$ é estimável, então, a soma de quadrados devida à hipótese H_0 pode ser obtida facilmente através da estatística de Wald:

$$\begin{aligned} SQH_0 &= (B'\hat{\mu})'[B'(W'W)^{-1}B]^{-1}(B'\hat{\mu}) \\ &= y'\{W(W'W)^{-1}B[B'(W'W)^{-1}B]^{-1}B'(W'W)^{-1}W'\}y = y'P_1y, \end{aligned}$$

onde $\hat{\mu} = (W'W)^{-1}W'y$ e $W'W = \text{diag}(n_{ij})$. Portanto, sob H_o , SQH_o / σ^2 tem distribuição qui-quadrado não central com s graus de liberdade. A soma de quadrados do Resíduo pode ser obtida por:

$$SQRes = y'[I - W(W'W)^{-1}W']y = y'(I - P)y ,$$

onde $y'(I - P)y / \sigma^2$ tem distribuição qui-quadrado central com $[n - r(W'W)]$ graus de liberdade. Uma estimativa não-viesada para σ^2 é dada por:

$$QMRes = \hat{\sigma}^2 = SQRes / [n - r(W'W)] = \{y'(I - P)y\} / [n - r(W'W)] .$$

Como $P_1(I - P) = \phi$ então, sob normalidade, as formas quadráticas $y'P_1y$ e $y'(I - P)y$ são independentes. Portanto, sob H_o ,

$$\frac{SQH_o / s}{[y'y - y'W(W'W)^{-1}W'y] / [n - r(W'W)]} \sim F_{\{s ; [n - r(W'W)]\}} .$$

Um outro conjunto de hipóteses de interesse dos pesquisadores é descrito pelas hipóteses do tipo I sobre as médias ponderadas não ajustadas, que correspondem ao método H_1 de Henderson, e são dadas por:

$$H_A^* : \left(\sum_{j=1}^3 n_{1j} \mu_{1j} \right) / \sum_{j=1}^3 n_{1j} = \left(\sum_{j=1}^3 n_{2j} \mu_{2j} \right) / \sum_{j=1}^3 n_{2j} ,$$

$$H_B^* : \left(\sum_{i=1}^2 n_{i1} \mu_{i1} \right) / \sum_{i=1}^2 n_{i1} = \left(\sum_{i=1}^2 n_{i2} \mu_{i2} \right) / \sum_{i=1}^2 n_{i2} = \left(\sum_{i=1}^2 n_{i3} \mu_{i3} \right) / \sum_{i=1}^2 n_{i3} ,$$

ou, equivalentemente,

$$H_A^* : (n_{11}\mu_{11} + n_{12}\mu_{12} + n_{13}\mu_{13}) / n_{1\bullet} = (n_{21}\mu_{21} + n_{22}\mu_{22} + n_{23}\mu_{23}) / n_{2\bullet} ,$$

$$H_B^* : (n_{11}\mu_{11} + n_{21}\mu_{21}) / n_{\bullet 1} = (n_{12}\mu_{12} + n_{22}\mu_{22}) / n_{\bullet 2} = (n_{13}\mu_{13} + n_{23}\mu_{23}) / n_{\bullet 3} .$$

O método OS₂ de Overall & Spiegel (1969) fornece hipóteses do tipo II sobre as médias ponderadas ajustadas, que são, sem dúvida, visivelmente mais complexas.

$$H_A^{**} : \sum_{j=1}^3 \left(n_{ij} - \frac{n_{ij}^2}{n_{\bullet j}} \right) \mu_{ij} - \sum_{i' \neq i}^2 \sum_{j=1}^3 \frac{n_{ij} n_{i'j}}{n_{\bullet j}} \mu_{i'j} = 0 ,$$

$$H_B^{**} : \sum_{i=1}^2 \left(n_{ij} - \frac{n_{ij}^2}{n_{i\bullet}} \right) \mu_{ij} - \sum_{j' \neq j}^3 \sum_{i=1}^2 \frac{n_{ij} n_{ij'}}{n_{i\bullet}} \mu_{ij'} = 0 .$$

Sendo assim, de acordo com a ordenação de interesse estabelecida *a priori*, as hipóteses H_A^* e H_B^{**} ou H_B^* e H_A^{**} correspondem às hipóteses testadas pelo método OS₃ de Overall & Spiegel (1969).

Finalmente, as hipóteses sobre a interação A*B, sugeridas nos três métodos de Overall & Spiegel são equivalentes e, têm a forma:

$$H_{AB} : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23} .$$

Carlson & Timm (1974) mostram também que as somas de quadrados obtidas através da redução nas somas de quadrados do resíduo, dependem do tipo de restrições não estimáveis que são impostas para a realização de testes de hipóteses dos efeitos principais. Assim, por exemplo, se forem usadas as restrições não estimáveis do

tipo $\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_i = \sum_j (\alpha\beta)_j = 0$, na reparametrização do modelo de posto

incompleto, obtendo-se o modelo reparametrizado de posto completo, as somas de quadrados dos efeitos principais e interação $A*B$ correspondem às hipóteses definidas anteriormente. Em resumo, tem-se:

<u>Hipóteses</u>	<u>Somas de quadrados do resíduo condicional correspondentes</u>
------------------	--

$$H_A \Rightarrow SQ_{Res}[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*] = SQ_{Res}[\mu^*, \beta^*, (\alpha\beta)^*] - SQ_{Res}[\mu^*, \alpha^*, \beta^*, (\alpha\beta)^*]$$

$$H_B \Rightarrow SQ_{Res}[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*] = SQ_{Res}[\mu^*, \alpha^*, (\alpha\beta)^*] - SQ_{Res}[\mu^*, \alpha^*, \beta^*, (\alpha\beta)^*]$$

$$H_A^* \Rightarrow SQ_{Res}[\alpha^* | \mu^*] = SQ_{Res}[\mu^*] - SQ_{Res}[\mu^*, \alpha^*]$$

$$H_B^* \Rightarrow SQ_{Res}[\beta^* | \mu^*] = SQ_{Res}[\mu^*] - SQ_{Res}[\mu^*, \beta^*]$$

$$H_A^{**} \Rightarrow SQ_{Res}[\alpha^* | \mu^*, \beta^*] = SQ_{Res}[\mu^*, \beta^*] - SQ_{Res}[\mu^*, \alpha^*, \beta^*]$$

$$H_B^{**} \Rightarrow SQ_{Res}[\beta^* | \mu^*, \alpha^*] = SQ_{Res}[\mu^*, \alpha^*] - SQ_{Res}[\mu^*, \alpha^*, \beta^*]$$

$$H_{AB} \Rightarrow SQ_{Res}[(\alpha\beta)^* | \mu^*, \alpha^*, \beta^*] = SQ_{Res}[\mu^*, \alpha^*, \beta^*] - SQ_{Res}[\mu^*, \alpha^*, \beta^*, (\alpha\beta)^*]$$

Por outro lado, se forem impostas as restrições não estimáveis do tipo

$$\sum_i v_i \alpha_i = \sum_j w_j \beta_j = \sum_i v_i (\alpha\beta)_i = \sum_j w_j (\alpha\beta)_j = 0, \text{ então, as somas de quadrados,}$$

$SQ_{Res}[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*]$ e $SQ_{Res}[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*]$, não correspondem às hipóteses

H_A e H_B . As outras reduções nas somas de quadrados do resíduo, $SQ_{Res}[\alpha^* | \mu^*]$,

$SQ_{Res}[\beta^* | \mu^*]$, $SQ_{Res}[\alpha^* | \mu^*, \beta^*]$ e $SQ_{Res}[\beta^* | \mu^*, \alpha^*]$, são equivalentes sob quaisquer

condições impostas e, portanto, testam as hipóteses H_A^* , H_B^* , H_A^{**} e H_B^{**} .

Segundo Speed & Hocking (1976), o método da redução nas somas de quadrados, denotada através da notação-R (.), é utilizada por muitos autores no cálculo de várias somas de quadrados do quadro de análise da variância para dados desbalanceados. Embora haja somente uma notação-R (.), existem duas formas distintas em que ela pode ser calculada. Tomam, como exemplo, um modelo com dois fatores cruzados (A, B) e interação,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2 ; j = 1, 2, 3 ; n_{11} = 1 \text{ e os demais } n_{ij} = 2 ,$$

e mostram que uma forma é o procedimento 1, usado por Searle (1971) em modelos de posto incompleto. A outra forma é o procedimento 2, usado por Harvey (1968), Overall & Klett (1972), Carlson & Timm (1974) em modelos reparametrizados de posto completo. Nesse artigo, eles enfatizam que esses procedimentos não são idênticos, e que um não é caso especial do outro. Eles ressaltam ainda que as somas de quadrados obtidas através da notação-R (.) nos dois procedimentos não especificam as hipóteses testadas, fazendo com que vários autores não interpretem corretamente as hipóteses que estão sendo testadas. Speed & Hocking (1976) também enfatizam que as somas de quadrados obtidas através do procedimento 1, $R[\alpha \mid \mu, \beta, (\alpha\beta)]$ e $R[\beta \mid \mu, \alpha, (\alpha\beta)]$ são iguais a zero e, portanto, não testam as hipóteses H_A e H_B . No procedimento 2, usando-se as restrições não estimáveis do tipo $\sum_i v_i \alpha_i = \sum_j w_j \beta_j = \sum_i v_i (\alpha\beta)_i = \sum_j w_j (\alpha\beta)_j = 0$, as somas de quadrados, $R[\alpha^* \mid \mu^*, \beta^*, (\alpha\beta)^*]$ e $R[\beta^* \mid \mu^*, \alpha^*, (\alpha\beta)^*]$, não correspondem

às hipóteses H_A e H_B , mas às hipóteses H_A^* e H_B^* . Entretanto, quando se usam as restrições não estimáveis do tipo $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$, as somas de quadrados referentes aos efeitos principais, $R[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*]$ e $R[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*]$, correspondem às hipóteses,

$$H_A^{***} : \mu_{i1} = \mu_{i'1} \quad \forall i, i' (i \neq i') ,$$

$$H_B^{***} : \mu_{1j} = \mu_{1j'} \quad \forall j, j' (j \neq j') .$$

Portanto, eles concluem que a notação-R (.) não deve ser usada para descrever quais as hipóteses testadas, pois ela é apenas um algoritmo eficiente na obtenção das somas de quadrados.

Segundo Hocking & Speed (1975), Harvey (1960) afirma que, embora o uso de restrições não estimáveis do tipo $\alpha_a = \beta_b = (\alpha\beta)_{ib} = (\alpha\beta)_{aj} = 0$, forneça solução única para o sistema de equações normais, infelizmente, as estimativas dos parâmetros são inteiramente insatisfatórias, e podem conduzir a erros graves, fornecendo somas de quadrados incorretas. Hocking & Speed (1975), entretanto, enfatizam que essas somas de quadrados não são incorretas, mas simplesmente elas correspondem a hipóteses diferentes que, de fato, podem não ser de interesse para os pesquisadores.

Realmente, a afirmação de Hocking & Speed (1975) é coerente, pois se forem usadas as restrições do tipo $\alpha_a = \beta_b = (\alpha\beta)_{ib} = (\alpha\beta)_{aj} = 0$, as somas de quadrados referentes aos efeitos principais, $R[\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*]$ e $R[\beta^* | \mu^*, \alpha^*, (\alpha\beta)^*]$, correspondem às hipóteses,

$$H_A^{****} : \mu_{ib} = \mu_{i'b} \quad \forall i, i' (i \neq i') ,$$

$$H_B^{****} : \mu_{aj} = \mu_{aj'} \quad \forall j, j' (j \neq j') .$$

Speed *et al.* (1978) fazem uma revisão dos principais métodos existentes na literatura para a análise de delineamentos experimentais com dados desbalanceados e os relacionam com os *softwares* estatísticos disponíveis. Tomando como exemplo, um modelo com dois fatores cruzados e interação, constataam que os métodos diferenciam de acordo com as hipóteses associadas às somas de quadrados fornecidas por estes métodos. Segundo os autores, a escolha de um método particular para a análise de delineamentos experimentais com dados desbalanceados, deveria ser baseada na adequação das hipóteses à natureza dos dados, em lugar das conveniências computacionais e/ou ortogonalidade das formas quadráticas. Descrevem, também, as somas de quadrados utilizando a notação-R (.) em modelos de posto incompleto, porém, as hipóteses são desenvolvidas através de médias de caselas.

Os autores enfatizam que, por exemplo, a hipótese H_A é de fácil interpretação, porque declara que não existe diferença entre os níveis do fator A, quando avaliado sobre todos os níveis do fator B. As hipóteses H_A^* e H_A^{**} não são fáceis de compreender porque representam comparações entre médias ponderadas, sendo que os pesos são funções complexas do número de observações nas caselas. A hipótese H_A^{***} é totalmente diferente das três primeiras, pois prevê somente comparações entre médias de caselas da primeira coluna. Embora a hipótese H_A^{****} seja de fácil compreensão, pode não ter interesse prático.

Urquhart & Weeks (1978) enfatizam que as estruturas de dados desbalanceados fornecem resultados que se alteram conforme a posição das caselas vazias e conforme a distribuição das observações sobre as caselas ocupadas. Sendo assim, eles sugerem que o pesquisador, talvez em colaboração com o estatístico, deveria definir as funções paramétricas relevantes através de médias de caselas. O modelo de médias de caselas possibilita ao pesquisador a flexibilidade de especificar precisamente as funções estimáveis de interesse.

Hocking *et al.* (1980) propõem uma alternativa para testar hipóteses com dados desbalanceados, quando o pesquisador defronta-se com caselas vazias, e apresentam um algoritmo para obter as hipóteses desejadas. Assim, a metodologia proposta inicia-se com aquelas hipóteses que deveriam ser testadas se o experimento fosse balanceado. Então, com base nessas hipóteses, determinam-se as hipóteses efetivas quando ocorrem caselas vazias. A fim de evitar as ambigüidades usuais dos modelos superparametrizados, eles descrevem a metodologia através de médias de caselas e apresentam exemplos para:

(i) - blocos incompletos balanceados, com o modelo:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

sujeito à restrição $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$, para $i, j = 1, 2, 3$ com $n_{11} = n_{22} = n_{33} = 0$,

(ii) - blocos casualizados, através do modelo:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

sujeito à restrição $\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'} = 0$; $i, j = 1, 2, 3$ com $n_{12} = n_{21} = n_{23} = n_{32} = 0$,

(iii) - modelos com três fatores cruzados e interações,

$$y_{ijkt} = \mu_{ijk} + \varepsilon_{ijkt}$$

sujeito à restrição $\mu_{ijk} - \mu_{ij'k} - \mu_{i'jk} + \mu_{i'j'k} = \mu_{ijk'} - \mu_{ij'k'} - \mu_{i'jk'} + \mu_{i'j'k'} = 0$; para $i = 1, 2$ e $j, k = 1, 2, 3$; com $n_{112} = n_{121} = n_{123} = n_{221} = n_{223} = 0$.

Goodnight (1980), usando o conceito de estimabilidade, desenvolve testes de hipóteses para modelos de efeitos fixos sem recorrer às suposições usuais, por exemplo, restrições do tipo $\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_i = \sum_j (\alpha\beta)_j = 0$, para reparametrizar o modelo de posto incompleto. Ele alega que as somas de quadrados obtidas através do uso de tais suposições, têm confundido os pesquisadores com relação às interpretações das hipóteses estatísticas. Isto porque existem diferentes tipos de restrições usuais que podem levar a hipóteses distintas. Tomando como exemplo um modelo com dois fatores cruzados (A, B) e interação,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$i = 1, 2$; $j = 1, 2$; $n_{22} = 2$ e os demais $n_{ij} = 1$,

define três tipos de funções estimáveis (I, II e III), sendo que o tipo I é usado para números diferentes de observações nas caselas; o tipo II em presença de caselas vazias e o tipo III na ocorrência de qualquer grau de confundimento dos efeitos em modelos lineares de efeitos fixos.

Searle (1980) mostra que o procedimento GLM do SAS inclui em sua saída quatro tipos de funções estimáveis arbitrárias representadas pela letra L em seus coeficientes. Considerando, como exemplo, um modelo com dois fatores cruzados (A, B) e interação,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2, \dots, a ; j = 1, 2, \dots, b ; k = 1, 2, \dots, n_{ij} ,$$

mostra como tais funções estimáveis arbitrárias são obtidas das expressões gerais para os testes de hipóteses através da estatística F, na análise de variância com dados desbalanceados. Segundo o autor, muitos usuários que trabalham com *softwares* estatísticos, primeiro, obtêm a estatística F na análise de variância, sem antes formular as hipóteses testadas. Sendo que, o procedimento correto para se efetuar testes de hipóteses consiste, na formulação, *a priori*, da hipótese. Só, então, deve-se testá-la através da estatística F.

Searle *et al.* (1981) enfatizam que há na literatura no mínimo 5 métodos diferentes para os cálculos das somas de quadrados na análise de variância com dados desbalanceados que estão incorporados pelos *softwares* estatísticos. Eles são caracterizados por: método para modelos com restrição do tipo Σ , que impõe restrições não estimá-

veis nos modelos superparametrizados, tornando-os de posto completo; método da inversa de parte inversa da matriz $X'X$ (tipo III do SAS-GLM); método dos quadrados de médias ponderadas (Yates, 1934); método adotado por Searle (1971), utilizando a notação-R (.) para modelos de posto incompleto e o método para testar hipótese de interesse da forma $H_0 : K'b = m$, utilizando a expressão SQH_0 de Wald.

Embora esses métodos produzam resultados idênticos para modelos com dados balanceados, eles produzem resultados diferentes quando os dados são desbalanceados. Além disso, para utilizar esses métodos, é necessário basear-se em funções estimáveis. Se, entretanto, existem caselas vazias, por exemplo, no modelo com dois fatores cruzados (A, B) e interação, diferenças entre os níveis dos fatores podem não ser estimáveis e, nesses casos, as hipóteses correspondentes não são testáveis. Uma alternativa para calcular as somas de quadrados em casos de caselas vazias pode ser obtida através do modelo de médias de caselas (modelo-M), que apresenta menos problemas de estimabilidade. O modelo-M simplifica as interpretações das hipóteses, além de permitir testar hipóteses baseadas em funções estimáveis.

Padovani (1984), utilizando o conceito clássico sobre as funções lineares paramétricas estimáveis, estabelece um teorema que resulta numa regra simples e prática para a verificação de funções estimáveis nos modelos de classificação hierárquica com s estágios, para dados desbalanceados. Com base na metodologia apresentada, desenvolve também, um programa computacional em FORTRAN IV para verificar se a função proposta pelo pesquisador é estimável e, caso seja, o algoritmo realiza testes de hipóteses

correspondentes. Finalizando, ilustra o desenvolvimento teórico com uma aplicação a dados de comprimentos de fibras de madeira.

Elliott & Woodward (1986) comparam cinco *softwares* estatísticos para microcomputadores, SPSS/PC (*Statistical Package for the Social Sciences*), SYSTAT (*The System for Statistics*), NCSS (*Number Cruncher Statistical System*), STATPAC desenvolvido por *Walonick Associates, inc.* e STATPAL desenvolvido por *Marcel Dekker, inc.*. Infelizmente, os autores não acataram as recomendações de Hamer (1981), onde ele enfatiza que, os artigos que avaliam programas de computadores devem incluir sempre, as versões e/ou atualizações em todos os casos, bem como utilizar programas atualizados. Utilizando como exemplo, um modelo com dois fatores cruzados (A, B) e interação, definidos através de médias de caselas,

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2, 3, 4 ; j = 1, 2, 3 ; k = 1, 2, \dots, n_{ij} ,$$

eles avaliam, para cada *software*, as respectivas somas de quadrados e hipóteses testadas sob várias opções dos programas. Também enfatizam que muitos *softwares* estatísticos para microcomputadores foram avaliados extensivamente pela comunidade estatística e os resultados publicados em revistas científicas. Segundo os autores, foi devido a artigos publicados, dessa natureza, que muitos algoritmos deficientes, foram corrigidos para fornecerem cálculos estatísticos corretos. Além disso, esses *softwares* estatísticos para microcomputadores estão sendo utilizados, provavelmente, mais por pesquisadores de

outras áreas com conhecimentos estatísticos limitados, que por estatísticos de formação. Segundo Elliott & Woodward (1986), como os *softwares* estatísticos estão se multiplicando, existem em torno de 200, e com um número estimado de usuários que excede 50.000, os estatísticos precisam avaliar melhor o que eles estão calculando. Essas avaliações permitem alertar os analistas de dados sobre o perigo de se utilizar *softwares* estatísticos sem os manuais bem documentados e sobre o que eles estão calculando sem os conhecimentos estatísticos necessários.

Iemma (1993 e 1995-c) apresenta, em detalhes, as hipóteses estatísticas mais utilizadas para testar os efeitos de linhas, colunas e interação no modelo com dois fatores cruzados (A, B) com e sem interação, para dados desbalanceados em presença de caselas vazias,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2 ; j = 1, 2, 3 ; n_{23} = 0 ; n_{12} = 1 ; n_{11} = n_{13} = n_{21} = 2 \text{ e } n_{22} = 3 .$$

Discute também o comportamento do SAS-GLM, versão 6.04, em relação aos métodos e hipóteses testadas.

Santos (1994) estabelece a estruturação de hipóteses estatísticas e somas de quadrados a elas associadas para dados desbalanceados. Visa, dessa forma, orientar a utilização de alguns *softwares* estatísticos por usuários não iniciados em estatística. Para tanto, o autor apresenta uma revisão dos métodos de análise de dados desbalanceados, adotando-se um modelo com dois fatores cruzados em presença ou não da interação,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2, 3 \text{ e } j = 1, 2, 3, 4 ; n_{12} = n_{23} = n_{31} = n_{34} = 0 ; n_{13} = n_{32} = 1 ;$$

$$n_{14} = n_{21} = n_{22} = n_{33} = 2 \text{ e } n_{11} = n_{24} = 3 .$$

Segundo o autor, quando se tem dados desbalanceados, as interpretações das verdadeiras hipóteses que são testadas através das somas de quadrados, obtidas por qualquer dos métodos disponíveis, podem trazer problemas. Assim, a estrutura de desbalanceamento pode induzir a hipóteses equivocadas, ou seja, pode-se imaginar estar testando uma hipótese $H_0^{(1)}$, quando na realidade está testando uma hipótese $H_0^{(2)}$, que não corresponde ao objetivo da pesquisa. Um outro agravante é o uso inadequado dos *softwares* estatísticos pois, em geral, os usuários não conhecem exatamente as hipóteses envolvidas. Finalmente, visando a uma orientação na análise e formulação das hipóteses, o autor faz uma comparação das saídas geradas pelos *softwares* estatísticos mais utilizados dentro das ciências agrárias, tais como, o SAS (versão 1982), o BMDP (versão 1983), o SYSTAT (*The System for Statistics* - versão 1986), o SOC (Sistema Operacional C - versão 1982), o STATGRAPHICS (versão 1985), o SANEST (versão 1986), o SAEG (versão 1986), o GLIM (versão 1984), o HARVEY (*Mixed Model Least-Squares and Maximum Likelihood Computer Program* - versão 1987), o MINITAB (versão 1986), o MSTAT (*A Microcomputer for the Design, Management and Analysis Agronomic Research Experiments* - versão 1984), e o SPSS (versão 1986).

Mondardo (1994), com o objetivo de alertar aos pesquisadores para a complexidade da análise de dados desbalanceados com caselas vazias e orientá-los na

escolha do método de análise mais adequado, através de dois exemplos com diferentes padrões de caselas vazias para o modelo com dois fatores cruzados e interação,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} ,$$

$$i = 1, 2 ; j = 1, 2, 3 ; n_{21} = 0 ; n_{11} = n_{12} = n_{23} = 1 ; n_{13} = n_{22} = 2 .$$

e

$$i = 1, 2, 3 ; j = 1, 2, 3, 4 ; n_{12} = n_{23} = n_{31} = n_{34} = 0 ;$$

$$n_{11} = n_{14} = n_{21} = n_{22} = n_{33} = 1 ; n_{13} = n_{24} = n_{32} = 2 ,$$

mostra passo a passo, a construção de funções estimáveis que geram os quatro tipos de somas de quadrados fornecidas pelo SAS-GLM, versão 6.04. Além da construção, apresenta tentativas no sentido de interpretar o que as somas de quadrados provenientes destas funções estão testando. Assim, de posse de tais informações, o pesquisador pode escolher quais dentre tais funções estimáveis são adequadas para representar a hipótese de seu interesse. Pode até mesmo concluir que nenhuma delas é adequada.

Segundo Searle (1994), Dallal (1992) apresenta um exemplo interessante mostrando algumas das dificuldades encontradas em interpretar as somas de quadrados produzidas por *softwares* estatísticos. É um dos primeiros trabalhos que tratam de modelos complexos com estrutura do tipo fatorial-hierárquica. Segundo o autor, Dallal (1992) relata tais dificuldades analisando os dados através de dois *softwares* SAS-GLM, versão 6.04 e SPSS-MANOVA, versão 4.0. Para cada um deles, são realizadas as análises usando dois tipos de modelos, porém estatisticamente similares. Os modelos são:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

(Modelo-S)

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{ij:k}$$

(Modelo-C)

O primeiro modelo considerado por Dallal (1992) é o tradicional modelo superparametrizado com dois fatores e interação. O segundo modelo que o autor utiliza é aquele no qual introduz um terceiro fator para classificar as observações dentro de cada casela, como fator C aninhado sob a interação A*B.

Tabela 1: Dados de Dallal

	B1	B2	B3
A1	3,81	3,42	3,55
	4,64	3,57	3,71
	4,09	3,55	3,66
A2	0,22	0,36	0,37
	0,33	0,27	0,31
	0,36	0,26	0,28
	1,08	0,83	0,70
	1,33	0,90	0,89
	1,15	0,93	0,93

As dificuldades encontradas em interpretar os resultados das somas de quadrados surgiram exatamente entre os Modelo-S e Modelo-C, onde as somas de quadrados do tipo III referentes ao fator B, fornecida pelo SAS-GLM, não são as mesmas para ambos modelos. Esperava-se, no entanto, que elas fossem as mesmas, pois à exceção de uma simples mudança de classificação, os dois modelos são estatisticamente

similares e, portanto, não deveria haver diferença nos cálculos das somas de quadrados da análise de variância.

Tabela 2: Somas de quadrados do tipo III

S.Q	Modelo-S	Modelo-C
SQA	59,11574074	59,11574074
SQB	0,78843333	<u>0,75733365</u>
SQ(A*B)	0,28218148	0,28218148
SQ[C(A*B)]	-	2,57130000
SQRes	2,57130000	-

Para tentar esclarecer a ocorrência dessa soma de quadrados do tipo III que não testa a hipótese do tipo III sobre as médias não ponderadas proposto por Yates (1934), quando se utiliza o SAS-GLM e o modelo de efeitos fixos hierarquizados em presença de esquema fatorial, Searle (1994) discute três razões.

A questão levantada por Searle (1994) foi, “porque o fator B é afetado, mas o fator A não?”. Parece não haver razão óbvia para um fator ser afetado de modo diferente em relação ao outro fator.

- Talvez uma das razões fosse a sequência de entrada dos fatores principais no modelo. Porém, a sequência na qual os fatores entram no modelo não faz diferença, pois essa sequência não altera a soma de quadrados do tipo III, fornecida pelo SAS-GLM, já que ela é ajustada para todos os fatores e interações.

- Para explicar uma segunda razão, observou os resultados das funções estimáveis do tipo III que geram as somas de quadrados do tipo III. Isto porque, segundo Searle

(1987), as funções estimáveis fornecem informações adequadas para a construção de hipóteses estatísticas testadas pelas somas de quadrados. Dessa forma, esperava-se que pudesse estabelecer uma explicação para a hipótese do tipo III e, portanto, compreender melhor o comportamento da soma de quadrados do tipo III, no modelo de efeitos fixos hierarquizados em presença de esquema fatorial. Infelizmente, isso não aconteceu. Segundo o autor, nos procedimentos da 1^a Conferência dos Usuários do SAS, Goodnight (1976) forneceu a descrição de como as funções estimáveis correspondentes às somas de quadrados do tipo III são derivadas, e é dessas funções estimáveis que as hipóteses podem ser construídas. Entretanto, todas as descrições dadas por Goodnight (1976) são apenas em termos de computação. Só existe um comentário estatístico na penúltima consideração, relativa ao método dos quadrados de médias ponderadas de Yates (1934), mas está restrita a dados com todas as caselas ocupadas e a experimentos fatoriais. Ele comenta: “a não ser que não exista uma explicação estatística detalhada sobre o que está sendo computado”. Desse modo, sem uma explicação estatística através dos manuais de utilização, como essa soma de quadrados do tipo III fornecida pelo SAS-GLM é calculada, e levando em consideração a hipótese que ela testa, tudo indica que nenhum procedimento estatístico usual é capaz de desenvolver aquela hipótese. Isto levou a considerar uma outra razão possível para o valor da soma de quadrados do tipo III referente ao fator B não ser a mesma nos dois modelos.

- O autor acredita que, essa diferença seja devida ao desbalanceamento dos dados com relação aos níveis do fator C(A*B). Por exemplo, observando-se os dados da Tabela 1, pode ser visto que, dentro de cada nível do fator A existe balanceamento com relação

aos níveis do fator $C(A*B)$ sobre os níveis do fator B. Assim, em A1 existem três níveis do fator $C(A*B)$ para cada nível do fator B, e em A2 há seis níveis do fator $C(A*B)$ para cada nível de B. Aparentemente, por causa desse aspecto de balanceamento, a soma de quadrados do tipo III referente ao fator A é a mesma para ambos os modelos e testa realmente a hipótese do tipo III. Ao contrário, essa forma de balanceamento não se aplica para a soma de quadrados do tipo III, referente ao fator B. Dentro de cada nível do fator B não existe balanceamento com relação aos níveis do fator $C(A*B)$ sobre os níveis do fator A. Assim, verifica-se que dentro do fator B existem três níveis do fator $C(A*B)$ em A1, mas seis níveis do fator $C(A*B)$ em A2. Aparentemente, isto fez com que a soma de quadrados do tipo III referente ao fator B no Modelo-C fosse diferente do Modelo-S e, portanto, não testasse a hipótese do tipo III.

Para verificar o argumento sobre desbalanceamento dos dados com relação aos níveis do fator $C(A*B)$, Searle (1994) retirou a observação y_{111} da Tabela 1 de Dallal, de modo que o fator A também ficasse desbalanceado. Nessa nova configuração, calculou novamente a soma de quadrados do tipo III. Constatou que agora as somas de quadrados do tipo III referentes aos fatores A e B, não testaram as hipóteses do tipo III, mas hipóteses completamente diferentes.

Segundo o autor, embora o raciocínio precedente sobre desbalanceamento dos níveis do fator $C(A*B)$ pareça fornecer uma explicação satisfatória para a ocorrência dessas diferenças nas somas de quadrados nos dois modelos, não fornece uma explicação estatística de como a soma de quadrados do tipo III referente ao fator B foi calculada no Modelo-C ou de como a hipótese foi derivada. Sendo assim, parece razoável não usar as

somas de quadrados do tipo III em modelos de efeitos fixos hierarquizados em presença de esquema fatorial. Finalizando, após analisar outros *softwares* estatísticos, tais como, o BMDP, o SPSS, o SYSTAT e o STATA, o autor conclui que as empresas fabricantes de *softwares* precisam fornecer através dos manuais, claramente, mais detalhes e, especialmente, mais descrições específicas sobre o que esses *softwares* estão calculando.

Camarinha Filho (1995) enfatiza que a falta de documentação em alguns *softwares* e a ausência de profissional especializado em análise de dados, pode levar a erros de interpretação, comprometendo os resultados obtidos a altos custos e vários anos de experimentação. Com o objetivo de amenizar essa deficiência e visando iniciar o pesquisador, principalmente das ciências aplicadas, nas idéias sobre as interpretações das hipóteses testadas, utiliza dois *softwares* estatísticos, o SAS-GLM, versão 6.04 e o STATGRAPHICS, versão 7.0, a fim de estudar os resultados das somas de quadrados e hipóteses associadas. Considerou um modelo com dois fatores cruzados (A, B) e interação e estudou diferentes padrões de caselas vazias.

Iemma (1997) enfatiza que a ocorrência de caselas vazias pode conduzir a graves problemas em relação à estimação de funções paramétricas e, portanto, à testabilidade de tais funções. Esses problemas refletem-se sobre as interpretações das verdadeiras hipóteses testadas pelas somas de quadrados fornecidas pelos *softwares* estatísticos, podendo causar sérios transtornos tanto aos estatísticos como aos pesquisadores das ciências aplicadas. Segundo o autor, alguns dos problemas mais importantes, motivados pela ocorrência de caselas vazias são dentre outros:

i) - Pode ocorrer desconexões;

- ii) - A hipótese sobre a interação não testa que todas as interações são nulas;
- iii) - As hipóteses sobre as médias ponderadas de linhas e de colunas têm, em geral, interpretação bastante diferentes daquelas obtidas quando todas as caselas estão ocupadas;
- iv) - A interpretação das hipóteses depende fortemente da posição e do número de caselas vazias;
- v) - Os graus de liberdade da interação devem ser escolhidos com cautela;
- vi) - O teste sobre a interação não deve ser utilizado como um critério de escolha entre um modelo “com” e um modelo “sem” interação.

3 METODOLOGIA

Seguindo a metodologia adotada por Scheffé (1959), Searle (1971), Hocking (1985), Searle (1987) e Iemma (1987 e 1993), são abordadas as hipóteses estatísticas e somas de quadrados a elas associadas, considerando-se os modelos de efeitos fixos com estruturas do tipo estritamente hierárquica e do tipo fatorial-hierárquica. Conforme visto no item 1, para esses modelos, são formulados e interpretados os termos das respectivas somas de quadrados, bem como as hipóteses estatísticas que são testadas para os efeitos principais e interações. Ademais, são discutidos os comportamentos de alguns *softwares* estatísticos mais utilizados dentro das ciências aplicadas em relação às somas de quadrados e hipóteses testadas.

3.1 Modelo Linear de Gauss-Markov Normal

Segundo Iemma (1987) entre outros, o modelo linear dado por:

$$y = X\theta + \varepsilon \quad (1)$$

onde, y é um vetor de realizações de variáveis aleatórias de dimensão $(n \times 1)$; X é uma matriz conhecida de dimensão $(n \times p)$ e de posto $k \leq \min\{n, p\}$; θ é um vetor de parâme-

tros desconhecidos, de dimensão $(p \times 1)$, sobre os quais concentram-se as hipóteses de interesse, e ε é um vetor de variáveis aleatórias não observáveis de dimensão $(n \times 1)$, tal que, $\varepsilon \sim N(\phi, I\sigma^2)$, é definido como modelo linear de Gauss-Markov Normal. Nesse caso, tem-se que a esperança matemática e a variância de y são dadas por $E[y] = X\theta$ e $V[y] = I\sigma^2$.

De acordo com Iemma (1993), o modelo linear de Gauss-Markov Normal pode assumir várias formas equivalentes. Entre outras citam-se: modelo superparametrizado (Modelo-S), modelo de médias de caselas (Modelo-M), modelo com restrição paramétrica do tipo sigma (Modelo- Σ) e modelo de regressão (Modelo-R). Segundo o autor, o modelo superparametrizado é excelente para a interpretação da notação-R (.) e de certas somas de quadrados a ela associadas. O modelo de médias de caselas simplifica a construção e a interpretação das hipóteses testadas. Os modelos com restrição- Σ e de regressão, facilitam a interpretação de certos tipos de somas de quadrados baseados em modelos de posto completo. Face aos objetivos deste trabalho, serão aqui considerados os modelos S e M.

3.2 Modelos com Estrutura do Tipo Estritamente Hierárquica

Nos experimentos fatoriais cruzados, cada nível de um fator ocorre com cada nível do outro fator, apresentando todas as combinações entre os níveis desses fatores, denotadas tratamentos. Já os experimentos com estrutura do tipo hierárquica diferem dos fatoriais por não apresentarem tais combinações, pois os níveis de um fator não são

os mesmos para cada nível dos outros fatores.

Visando evitar generalizações complexas, maçantes e desnecessárias, apresenta-se aqui o modelo, com base num experimento genérico com três fatores. Assim, como exemplo, seja um experimento 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

Para distinguir entre os experimentos fatoriais cruzados e experimentos com estrutura do tipo estritamente hierárquica, deve-se verificar:

- Se nenhum fator é hierarquizado sobre o outro. Nesse caso, os fatores são cruzados e cada nível de um fator ocorre com cada nível do(s) outro(s) fator(es).
- No entanto, se os fatores são hierarquizados, então, os níveis dos fatores podem ser enumerados arbitrariamente, já que os b_i níveis do fator B são diferentes em cada nível do fator A e c_{ij} níveis do fator C são diferentes em cada combinações entre

níveis dos fatores A e B.

Os modelos com estrutura do tipo estritamente hierárquica apresentam uma vasta aplicação em experimentos hospitalares, industriais, de controle estatístico de qualidade, entre outras. Também são utilizados constantemente em pesquisas com animais e vegetais, principalmente, nas áreas de melhoramento genético de bovinos de corte, suínos, aves e de plantas.

3.2.1 Modelo de Médias de Caselas (Modelo-M)

O modelo de médias de caselas adotado para as classificações hierárquicas com três fatores pode ser descrito na forma matricial como:

$$y = W\mu + \varepsilon \quad (2)$$

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

O modelo descrito na forma matricial em (2) pode ser caracterizado por:

$$y_{ijkt} = \mu_{ijk} + \varepsilon_{ijkt} \quad (3)$$

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

onde, y_{ijkt} é a resposta observada na t-ésima parcela que recebeu o k-ésimo nível do fator C aninhado sob o j-ésimo nível do fator B, sendo que o j-ésimo nível do fator B está aninhado sob o i-ésimo nível do fator A e, o i-ésimo nível do fator A; 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_{ijkt} , tal que $\varepsilon_{ijkt} \sim \text{Niid}(0, \sigma^2)$.

Definido o modelo em (2), o passo seguinte é obter um estimador para o parâmetro μ com base nas propriedades ótimas, tais como estimadores não-viesados, consistentes, eficientes e de variância mínima. Na literatura existem disponíveis vários métodos de estimação que satisfazem a maioria dessas propriedades. Dentre esses métodos, aquele que se destaca dentro da teoria de modelos lineares é o método de mínimos quadrados. O método de estimação por mínimos quadrados ordinários (OLSE - *Ordinary Least Squares Estimation*), que assume a estrutura de covariância da forma $I\sigma^2$, consiste em encontrar o valor do parâmetro μ que minimiza o vetor de erros, $\varepsilon'\varepsilon = (y - W\mu)'(y - W\mu)$, obtendo-se o sistema de equações normais (SEN),

$$W'W\mu = W'y \quad (4)$$

onde,

$$W'W = \text{diag}\{n_{ijk}\} \text{ e } W'y =$$

$$= [y_{111} \dots y_{11c_{11}} \dots \dots y_{1b_1 1} \dots y_{1b_1 c_{1b_1}} \dots \dots \dots y_{a11} \dots y_{a1c_{a1}} \dots \dots y_{ab_a 1} \dots y_{ab_a c_{ab_a}}]'$$

Como já mencionado, além de facilitar a construção e a interpretação das hipóteses estatísticas, o modelo de médias de caselas proporciona solução única para o vetor de médias, μ , pois a matriz W tem posto coluna completo. Sendo assim, a matriz $W'W$ é positiva definida e, portanto, existe a inversa clássica, $(W'W)^{-1} = \text{diag}\{1 / n_{ijk}\}$.

Nesse contexto, a solução exata do SEN, que é solução aproximada de mínimos quadrados para o sistema inconsistente $y = W\mu$, coincide com o estimador de mínimos quadrados,

$$\hat{\mu} = (W'W)^{-1} W'y \quad (5)$$

Sendo assim, o melhor estimador linear não-viesado (BLUE - *Best Linear Unbiased Estimator*) de $B'\mu$ estimável e sua variância, são dados por:

$$\text{BLUE } (B'\mu) = B'\hat{\mu} = B'(W'W)^{-1} W'y \quad (6)$$

$$\text{Var } [\text{BLUE } (B'\mu)] = \text{Var } (B'\hat{\mu}) = B'(W'W)^{-1} B\sigma^2 \quad (7)$$

3.2.2 Modelo Superparametrizado (Modelo-S)

Um modelo alternativo equivalente ao Modelo-M definido em (2) é aquele denominado Modelo-S ou modelo superparametrizado. Segundo Searle (1987), o modelo é superparametrizado quando apresenta explicitamente um parâmetro para cada efeito dos fatores envolvidos. Por exemplo, o procedimento GLM do SAS trabalha com o

modelo superparametrizado, tanto na apresentação de funções estimáveis quanto para as estimativas de cada um dos parâmetros do modelo linear.

De acordo com Iemma & Perri (1996), entre outros, os modelos superparametrizados são parte integrante da história dos modelos lineares e têm sido de grande valia para os pesquisadores das ciências aplicadas, pois eles mostram explicitamente os parâmetros sobre os quais concentram-se as hipóteses de interesse. Apresenta, no entanto, o inconveniente de possuir mais parâmetros que médias de caselas para estimá-los.

Assim, no modelo com três fatores hierarquizados, seguindo a caracterização adotada por Winer (1971), Anderson & Mclean (1974) dentre outros, tem-se:

$$y_{ijkt} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \varepsilon_{ijkt} \quad (8)$$

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

onde, y_{ijkt} e ε_{ijkt} , são como definidos em (3), 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 caracterizado em (8) na forma matricial, tem-se:

$$y = X\theta + \varepsilon , \quad \varepsilon \sim N(\phi, I\sigma^2). \quad (9)$$

Nesse caso, o sistema de equações normais é dado por:

$$X'X\theta = X'y \quad (10)$$

Segundo Iemma (1987), como X é de posto incompleto, a menos de reparametrizações, não existe a inversa clássica $(X'X)^{-1}$ e, portanto, o SEN é indeterminado. Uma solução dentre outras é dada por:

$$\theta^o = (X'X)^- X'y \quad (11)$$

onde, $(X'X)^-$ é uma inversa generalizada qualquer de $X'X$. Nesse caso, θ^o , solução exata do SEN, é solução aproximada de mínimos quadrados para o sistema inconsistente $y = X\theta$. Portanto, o BLUE de $\lambda'\theta$ estimável e sua variância são dados, através do teorema de Gauss-Markov, por:

$$\text{BLUE } (\lambda'\theta) = \lambda'\theta^o = \lambda'(X'X)^- X'y \quad (12)$$

$$\text{Var } [\text{BLUE } (\lambda'\theta)] = \text{Var } (\lambda'\theta^o) = \lambda'(X'X)^- X'X(X'X)^- \lambda\sigma^2, \quad (13)$$

invariante para qualquer θ^o , solução das equações normais e para qualquer escolha de $(X'X)^-$.

3.2.3 Obtenção das Somas de Quadrados

Segundo Iemma (1993), conforme o interesse do pesquisador, o modelo de médias de caselas e o modelo superparametrizado podem assumir diversas caracteriza-

ções, desde o modelo mais simples até modelos mais parametrizados. No modelo de médias, tem-se;

$$y = W_1\mu_1 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \varepsilon_{ijkt} \quad (\text{M.11})$$

$$y = W_2\mu_2 + \varepsilon \Leftrightarrow y_{ijkt} = \mu_i + \varepsilon_{ijkt} \quad (\text{M.12})$$

$$y = W_3\mu_3 + \varepsilon \Leftrightarrow y_{ijkt} = \mu_{ij} + \varepsilon_{ijkt} \quad (\text{M.13})$$

$$y = W\mu + \varepsilon \Leftrightarrow y_{ijkt} = \mu_{ijk} + \varepsilon_{ijkt} \quad (\text{M.14})$$

e no modelo superparametrizado,

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

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

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

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

Tais parametrizações sucessivas e ordenadas facilitam a interpretação da notação-R (.) e de certas somas de quadrados a elas associadas.

Dado o modelo linear $y = X\theta + \varepsilon$, $\varepsilon \sim N(\phi, I\sigma^2)$, então:

(i) - $\hat{y} = Py = X(X'X)^{-1}X'y = X\theta^0$, invariante para qualquer solução das equações normais, $X'X\theta = X'y$, é a aproximação de mínimos quadrados para $y \in \mathbb{R}^n$ e P é o projetor ortogonal de y sobre o espaço gerado pelas colunas de X , denotado $C(X)$.

(ii) - $\hat{\varepsilon} = [I - X(X'X)^{-1}X']y = (I - P)y = y - \hat{y}$ é o erro de ajuste devido à aproximação de mínimos quadrados para o vetor y , das observações e $(I - P)$ é o projetor ortogonal de y sobre o complemento ortogonal do espaço coluna de X , $C^\perp(X)$.

(iii) - $y = \hat{y} + \hat{\varepsilon}$, $y \in \mathbb{R}^n$, $\hat{y} \in C(X)$ e $\hat{\varepsilon} \in C^\perp(X)$.

Figura 2: Projeção ortogonal de y sobre $C(X)$

Dada a ortogonalidade de $\hat{y}$ e $\hat{\varepsilon}$, pode-se utilizar o teorema de Pitágoras e obter-se a decomposição ortogonal clássica da análise de variância:

$$\|y\|^2 = \|\hat{y}\|^2 + \|\hat{\varepsilon}\|^2$$

onde,

$$\|y\|^2 = y'y = \text{SQTotal}$$

$$\|\hat{y}\|^2 = \hat{y}'\hat{y} = y'Py = \theta^0'X'y = \text{SQParâmetros} = R(\theta)$$

$$\|\hat{\varepsilon}\|^2 = y'(I - P)y = y'y - \theta^0'X'y = \text{SQRes}$$

Assim, através de parametrizações sucessivas e ordenadas, pode-se obter

as somas de quadrados de parâmetros para cada um dos modelos particionados.

$$y = W_1\mu_1 + \varepsilon \Rightarrow W_1'W_1\mu_1 = W_1'y \Rightarrow R(\mu) = \hat{\mu}_1'W_1'y \quad (M.21)$$

$$y = W_2\mu_2 + \varepsilon \Rightarrow W_2'W_2\mu_2 = W_2'y \Rightarrow R(\mu_i) = \hat{\mu}_2'W_2'y \quad (M.22)$$

$$y = W_3\mu_3 + \varepsilon \Rightarrow W_3'W_3\mu_3 = W_3'y \Rightarrow R(\mu_{ij}) = \hat{\mu}_3'W_3'y \quad (M.23)$$

$$y = W\mu + \varepsilon \Rightarrow W'W\mu = W'y \Rightarrow R(\mu_{ijk}) = \hat{\mu}'W'y \quad (M.24)$$

$$y = X_1\theta_1 + \varepsilon \Rightarrow X_1'X_1\theta_1 = X_1'y \Rightarrow R(\mu) = \theta_1^\circ X_1'y \quad (S.21)$$

$$y = X_2\theta_2 + \varepsilon \Rightarrow X_2'X_2\theta_2 = X_2'y \Rightarrow R(\mu, \alpha) = \theta_2^\circ X_2'y \quad (S.22)$$

$$y = X_3\theta_3 + \varepsilon \Rightarrow X_3'X_3\theta_3 = X_3'y \Rightarrow R[\mu, \alpha, \beta(\alpha)] = \theta_3^\circ X_3'y \quad (S.23)$$

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

Procedendo-se aos ajustes, de acordo com o interesse, por exemplo, do tipo seqüencial que o procedimento GLM do SAS denota em sua saída por tipo I, têm-se as seguintes somas de quadrados ajustadas:

$$R(\mu_i | \mu) = R(\mu_i) - R(\mu) \equiv R(\alpha | \mu) = R(\mu, \alpha) - R(\mu) \quad (14)$$

$$\begin{aligned} R(\mu_{ij} | \mu_i) &= R(\mu_{ij}) - R(\mu_i) \equiv R[\beta(\alpha) | \mu, \alpha] \\ &= R[\mu, \alpha, \beta(\alpha)] - R(\mu, \alpha) \end{aligned} \quad (15)$$

$$\begin{aligned} R(\mu_{ijk} | \mu_{ij}) &= R(\mu_{ijk}) - R(\mu_{ij}) \equiv R[\gamma(\alpha\beta) | \mu, \alpha, \beta(\alpha)] \\ &= R[\mu, \alpha, \beta(\alpha), \gamma(\alpha\beta)] - R[\mu, \alpha, \beta(\alpha)] \end{aligned} \quad (16)$$

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

$$SQCorreção = n_{...} \bar{y}_{...}^2 \equiv R(\mu) \quad (17)$$

$$SQA = \sum_{i=1}^a n_{i..} (\bar{y}_{i...} - \bar{y}_{...})^2 \equiv R(\alpha | \mu) \quad (18)$$

$$SQB(A) = \sum_{i=1}^a \sum_{j=1}^{b_i} n_{ij.} (\bar{y}_{ij..} - \bar{y}_{i...})^2 \equiv R[\beta(\alpha) | \mu, \alpha] \quad (19)$$

$$SQC(A B) = \sum_{i=1}^a \sum_{j=1}^{b_i} \sum_{k=1}^{c_{ij}} n_{ijk.} (\bar{y}_{ijk.} - \bar{y}_{ij..})^2 \equiv R[\gamma(\alpha \beta) | \mu, \beta(\alpha)] \quad (20)$$

$$SQRes = \sum_{i=1}^a \sum_{j=1}^{b_i} \sum_{k=1}^{c_{ij}} \sum_{t=1}^{n_{ijk}} (y_{ijk.t} - \bar{y}_{ijk.})^2 \quad (21)$$

Como já enfatizado, as maiores dificuldades encontradas nas interpretações das hipóteses estatísticas testadas pelas somas de quadrados utilizando-se, tanto a notação-R (.) como os procedimentos usuais, é que elas não especificam as hipóteses testadas. Outro fato que deve ser considerado é que, embora o procedimento GLM do SAS denota (14), (15) e (16) como somas de quadrados do tipo I em sua saída e, de fato, no modelo com três fatores hierarquizados, elas realmente correspondem às hipóteses do tipo I sobre as médias ponderadas não ajustadas, em geral, isso não ocorre com outros modelos, por exemplo, em modelos com estrutura do tipo fatorial-hierárquica.

Searle (1971), Carlson & Timm (1974), Iemma (1987) entre outros, adotam um procedimento alternativo para obter as somas de quadrados baseadas em

hipóteses de interesse. Assim, para testar a hipótese do tipo $H_0 : B'\mu = \phi$, onde $B'\mu$ é um conjunto de funções estimáveis, B' tem posto linha completo e ϕ é um vetor de “zeros”, obtém-se a estatística de Wald:

$$SQH_0 = (B'\hat{\mu} - B\mu)'[B'(W'W)^{-1}B]^{-1}(B'\hat{\mu} - B'\mu) \quad (22)$$

onde,

$$\frac{SQH_0 / r(B)}{SQRes / [n - r(W'W)]} \sim F_{[r(B); n-r(W'W)]} \quad (23)$$

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

3.3 Modelos com Estrutura do Tipo Fatorial-Hierárquica

Em geral, é muito comum considerar modelos de classificações envolvendo vários fatores, onde esses fatores são todos cruzados ou hierarquizados. Entretanto, existem situações em que alguns fatores são cruzados, enquanto outros são hierarquizados. Experimentos com esse tipo de estrutura são denominados, segundo Anderson & Mclean (1974), experimentos com estrutura do tipo fatorial-hierárquica ou experimentos parcialmente hierarquizados (Lindman, 1974). Algumas vezes, tais modelos são utilizados na análise de experimentos com dois fatores e medidas repetidas sob o fator A. É interes-

sante ressaltar, também, que os modelos parcialmente hierarquizados têm muito em comum com o modelo adotado para experimentos com parcelas subdivididas, nome derivado pelo seu uso em experimentos agrícolas (Winer, 1971).

Figura 3: Esquema para Modelos com estrutura do tipo fatorial-hierárquica

Como os fatores A e B são cruzados, então, cada nível do fator B ocorre com cada nível do fator A, formando todas as combinações entre níveis desses fatores como tratamentos. Agora, como o fator C está aninhado sob os níveis do fator B, então, os níveis do fator C podem ser enumerados arbitrariamente, pois os c_j níveis do fator C são diferentes em cada nível do fator B.

3.3.1 - Modelo de Médias de Caselas (Modelo-M)

O modelo de médias de caselas adotado para a estrutura do tipo fatorial-hierárquica, conforme consta na Figura 3, difere daquele já visto para a estrutura do tipo estritamente hierárquica, pois aqui o fator A está cruzado com o fator B e o fator C está aninhado em cada nível do fator B. Esse tem sido um dos problemas do uso prático do

modelo-M, pois muitas vezes ele não evidencia diferenças existentes no plano experimental. Aqui como em (3) tem-se:

$$y_{ijkt} = \mu_{ijk} + \varepsilon_{ijkt} \quad (24)$$

$$i = 1, \dots, a ; j = 1, \dots, b ; k = 1, \dots, c_j ; t = 1, \dots, n_{ijk}$$

onde, y_{ijkt} é a resposta observada na t-ésima parcela que recebeu o k-ésimo nível do fator C aninhado sob o j-ésimo nível do fator B e o i-ésimo nível do fator A; μ_{ijk} é a média da população da qual foi retirada a amostra que compõe a casela (i, j, k) e ε_{ijkt} é o erro aleatório atribuído à observação, y_{ijkt} , tal que $\varepsilon_{ijkt} \sim \text{Niid}(0, \sigma^2)$.

O modelo caracterizado em (24) pode ser descrito na forma matricial do mesmo modo que em (2). Logo, o sistema de equações normais e sua solução exata, que coincide com o estimador de mínimos quadrados são equivalentes a (4) e (5). O BLUE de $B'\mu$ estimável e sua variância são equivalente a (6) e (7).

3.3.2 Modelo Superparametrizado (Modelo-S)

O modelo para a estrutura do tipo fatorial-hierárquica pode ser representado em termos de modelo superparametrizado, definindo-se um parâmetro para cada efeito considerado. Para a situação descrita na Figura 3, tem-se:

$$y_{ijkt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(j)} + (\alpha\gamma)_{ik(j)} + \varepsilon_{ijkt} \quad (25)$$

$$i = 1, \dots, a ; j = 1, \dots, b ; k = 1, \dots, c_j ; t = 1, \dots, n_{ijk}$$

onde, y_{ijkt} e ε_{ijkt} são como definidos em (24); α_i é o efeito do i -ésimo nível do fator A; β_j é o efeito do j -ésimo nível do fator B; $\gamma_{k(j)}$ é o efeito do k -ésimo nível do fator C aninhado sob o j -ésimo nível do fator B; $(\alpha\beta)_{ij}$ é a interação entre o i -ésimo nível do fator A com o j -ésimo nível do fator B e $(\alpha\gamma)_{ik(j)}$ é a interação entre o i -ésimo nível do fator A com o k -ésimo nível do fator C aninhado sob o j -ésimo nível do fator B.

Segundo Lindman (1974), o efeito principal do fator C, $\gamma_{k(j)}$, está confundido com o efeito da interação $(\beta\gamma)_{jk}$ e o efeito da interação $(\alpha\gamma)_{ik(j)}$ está confundido com o efeito da interação $(\alpha\beta\gamma)_{ijk}$.

A forma matricial do modelo caracterizado em (25) é semelhante ao definido em (9) e, portanto, o SEN e uma solução dentre outras, podem ser obtidas de (10) e (11), respectivamente. O BLUE de $\lambda'\theta$ estimável e sua variância são obtidas através de (12) e (13).

3.3.3 Obtenção das Somas de Quadrados

O procedimento utilizado para a obtenção das somas de quadrados no modelo com estrutura do tipo fatorial-hierárquica é semelhante àquele para o modelo com estrutura do tipo estritamente hierárquica. Assim, fazendo-se as parametrizações

sucessivas e ordenadas conforme o interesse, pode-se obter as somas de quadrados de parâmetros para cada um dos modelos particionados. No modelo de médias, tem-se:

$$y = W_1\mu_1 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \varepsilon_{ijkt} \quad (M.31)$$

$$y = W_2\mu_2 + \varepsilon \Leftrightarrow y_{ijkt} = \mu_i + \varepsilon_{ijkt} \quad (M.32)$$

$$y = W_3\mu_3 + \varepsilon \Leftrightarrow y_{ijkt} = \mu_i + \beta_j + \varepsilon_{ijkt} \quad (M.33)$$

$$y = W_4\mu_4 + \varepsilon \Leftrightarrow y_{ijkt} = \mu_{ij} + \varepsilon_{ijkt} \quad (M.34)$$

$$y = W_5\mu_5 + \varepsilon \Leftrightarrow y_{ijkt} = \mu_{ij} + \gamma_{k(j)} + \varepsilon_{ijkt} \quad (M.35)$$

$$y = W\mu + \varepsilon \Leftrightarrow y_{ijkt} = \mu_{ijk} + \varepsilon_{ijkt} \quad (M.36)$$

e, no modelo superparametrizado:

$$y = X_1\theta_1 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \varepsilon_{ijkt} \quad (S.31)$$

$$y = X_2\theta_2 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \alpha_i + \varepsilon_{ijkt} \quad (S.32)$$

$$y = X_3\theta_3 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \alpha_i + \beta_j + \varepsilon_{ijkt} \quad (S.33)$$

$$y = X_4\theta_4 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijkt} \quad (S.34)$$

$$y = X_5\theta_5 + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(j)} + \varepsilon_{ijkt} \quad (S.35)$$

$$y = X\theta + \varepsilon \Leftrightarrow y_{ijkt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(j)} + (\alpha\gamma)_{ik(j)} + \varepsilon_{ijkt} \quad (S.36)$$

Essas parametrizações sucessivas e ordenadas, facilitam a obtenção das somas de quadrados de parâmetros para cada um dos modelos.

$$y = W_1\mu_1 + \varepsilon \Rightarrow W_1'W_1\mu_1 = W_1'y \Rightarrow R(\mu) = \hat{\mu}_1'W_1'y \quad (M.41)$$

$$y = W_2\mu_2 + \varepsilon \Rightarrow W_2'W_2\mu_2 = W_2'y \Rightarrow R(\mu_i) = \hat{\mu}_2'W_2'y \quad (M.42)$$

$$y = W_3\mu_3 + \varepsilon \Rightarrow W_3'W_3\mu_3 = W_3'y \Rightarrow R(\mu_i + \beta_j) = \mu_3^0'W_3'y \quad (M.43)$$

$$y = W_4\mu_4 + \varepsilon \Rightarrow W_4'W_4\mu_4 = W_4'y \Rightarrow R(\mu_{ij}) = \hat{\mu}_4'W_4'y \quad (M.44)$$

$$y = W_5\mu_5 + \varepsilon \Rightarrow W_5'W_5\mu_5 = W_5'y \Rightarrow R(\mu_{ij} + \gamma_{k(j)}) = \mu_5^0'W_5'y \quad (M.45)$$

$$y = W\mu + \varepsilon \Rightarrow W'W\mu = W'y \Rightarrow R(\mu_{ijk}) = \hat{\mu}'W'y \quad (M.46)$$

$$y = X_1\theta_1 + \varepsilon \Rightarrow X_1'X_1\theta_1 = X_1'y \Rightarrow R(\mu) = \theta_1^0'X_1'y \quad (S.41)$$

$$y = X_2\theta_2 + \varepsilon \Rightarrow X_2'X_2\theta_2 = X_2'y \Rightarrow R(\mu, \alpha) = \theta_2^0'X_2'y \quad (S.42)$$

$$y = X_3\theta_3 + \varepsilon \Rightarrow X_3'X_3\theta_3 = X_3'y \Rightarrow R(\mu, \alpha, \beta) = \theta_3^0'X_3'y \quad (S.43)$$

$$y = X_4\theta_4 + \varepsilon \Rightarrow X_4'X_4\theta_4 = X_4'y \Rightarrow R[\mu, \alpha, \beta, (\alpha\beta)] = \theta_4^0'X_4'y \quad (S.44)$$

$$y = X_5\theta_5 + \varepsilon \Rightarrow X_5'X_5\theta_5 = X_5'y \Rightarrow R[\mu, \alpha, \beta, (\alpha\beta), \gamma] = \theta_5^0'X_5'y \quad (S.45)$$

$$y = X\theta + \varepsilon \Rightarrow X'X\theta = X'y \Rightarrow R[\mu, \alpha, (\alpha\beta), \gamma, (\alpha\gamma)] = \theta^0'X'y \quad (S.46)$$

Novamente, fazendo-se os ajustes do tipo seqüencial, obtém-se as somas de quadrados ajustadas:

$$R(\mu_i | \mu) = R(\mu_i) - R(\mu) \equiv R(\alpha | \mu) = R(\mu, \alpha) - R(\mu) \quad (26)$$

$$\begin{aligned} R(\mu_i + \beta_j | \mu_i) &= R(\mu_i + \beta_j) - R(\mu_i) \equiv R(\beta | \mu, \alpha) \\ &= R(\mu, \alpha, \beta) - R(\mu, \alpha) \end{aligned} \quad (27)$$

$$\begin{aligned} R(\mu_{ij} | \mu_i + \beta_j) &= R(\mu_{ij}) - R(\mu_i + \beta_j) \\ &\equiv R[(\alpha\beta) | \mu, \alpha, \beta] - R(\mu, \alpha, \beta) \end{aligned} \quad (28)$$

$$\begin{aligned} R(\mu_{ij} + \gamma_{k(j)} | \mu_{ij}) &= R(\mu_{ij} + \gamma_{k(j)}) - R(\mu_{ij}) \equiv R(\gamma | \mu, \alpha, \beta, (\alpha\beta)) \\ &= R[\mu, \alpha, \beta, (\alpha\beta), \gamma] - R[\mu, \alpha, \beta, (\alpha\beta)] \end{aligned} \quad (29)$$

$$\begin{aligned} R(\mu_{ijk} | \mu_{ij} + \gamma_{k(j)}) &= R(\mu_{ijk}) - R(\mu_{ij} + \gamma_{k(j)}) \equiv R[(\alpha\gamma) | \mu, \alpha, \beta, (\alpha\beta), \gamma] \\ &= R[\mu, \alpha, \beta, (\alpha\beta), \gamma, (\alpha\gamma)] - R[\mu, \alpha, \beta, (\alpha\beta), \gamma] \end{aligned} \quad (30)$$

Com a finalidade de comparar as hipóteses, consideram-se os procedimentos usuais, dados por:

$$SQCorreção = n_{...} \bar{y}_{...}^2 \equiv R(\mu) \quad (31)$$

$$SQA = \sum_{i=1}^a n_{i..} (\bar{y}_{i..} - \bar{y}_{...})^2 \equiv R(\alpha | \mu) \quad (32)$$

$$SQB = \sum_{j=1}^b n_{.j.} (\bar{y}_{.j.} - \bar{y}_{...})^2 \equiv R(\beta | \mu) \quad (33)$$

$$\begin{aligned} SQ(A*B) &= \sum_{i=1}^a \sum_{j=1}^b n_{ij.} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\ &\equiv R[\mu, \alpha, \beta, (\alpha\beta)] - R(\mu, \alpha) - R(\mu, \beta) + R(\mu) \end{aligned} \quad (34)$$

$$SQC(B) = \sum_{j=1}^b \sum_{k=1}^{c_j} n_{.jk.} (\bar{y}_{.jk.} - \bar{y}_{.j.})^2 \equiv R(\gamma | \mu, \beta) \quad (35)$$

$$\begin{aligned} SQ(A*C(B)) &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{c_j} n_{ijk.} (\bar{y}_{ijk.} - \bar{y}_{ij.} - \bar{y}_{.jk.} + \bar{y}_{.j.})^2 \\ &\equiv R[\mu, \alpha, \beta, (\alpha\beta), \gamma, (\alpha\gamma)] - R[\mu, \alpha, \beta, (\alpha\beta)] - R(\mu, \beta, \gamma) + R(\mu, \beta) \end{aligned} \quad (36)$$

$$SQRes = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{c_j} \sum_{t=1}^{n_{ijk}} (y_{ijkt} - \bar{y}_{ijk\bullet})^2 \quad (37)$$

Como já enfatizado no modelo com estrutura do tipo estritamente hierárquica, aqui também há dificuldades nas interpretações das hipóteses, pois tanto a notação-R (.) como os procedimentos usuais não especificam as hipóteses testadas.

Além disso, existem outras formas de parametrizações, por exemplo, ajustes dos tipos II e III do SAS-GLM que podem ser obtidos através da notação-R (.). Nesse caso, se há desbalanceamento em presença de caselas vazias, as somas de quadrados obtidas através da notação-R (.) podem testar hipóteses não usuais, muitas vezes sem interesse prático para os usuários.

Ademais, nos modelos com estrutura do tipo fatorial-hierárquica, se os dados são desbalanceados em presença ou não de caselas vazias, as somas de quadrados obtidas através dos procedimentos usuais são diferentes das somas de quadrados seqüenciais obtidas através da notação-R (.) e, portanto, testam hipóteses distintas. Isto porque, das somas de quadrados definidas em (32)-(36) através dos procedimentos usuais, apenas (32), (33) e (35) estão associadas às hipóteses do tipo I sobre as médias ponderadas não ajustadas. Do mesmo modo, das somas de quadrados obtidas através da notação-R (.), e que o SAS-GLM denota por tipo I em sua saída, apenas a soma de quadrados definida em (26) testa hipótese correspondente à do tipo I.

3.4 Hipóteses Estatísticas Usualmente Testadas Pelo SAS-GLM

Usando os modelos de efeitos fixos com estruturas do tipo estritamente hierárquica e do tipo fatorial-hierárquica, são formalizadas e discutidas as hipóteses mais usuais sobre os efeitos principais e interações. Deve-se ressaltar que, nesse estudo, as formulações de tais hipóteses são baseadas em médias de caselas e associadas às somas de quadrados fornecidas pelo procedimento GLM do SAS. Tal estratégia pode ser justificada pelo fato de que o SAS-GLM é largamente utilizado dentro das ciências aplicadas devido à vasta gama de opções que esse *software* oferece. Os usuários, entretanto, devem estar atentos, pois se os dados são desbalanceados em presença ou não de caselas vazias, as hipóteses testadas pelo SAS-GLM podem tornar-se extremamente difíceis de interpretar, principalmente em delineamentos complexos envolvendo 3 ou mais fatores. Por exemplo, o SAS-GLM fornece em sua saída quatro tipos de somas de quadrados: I, II, III e IV, muitas vezes geradas a partir de funções estimáveis complexas e, portanto, testam hipóteses em geral complexas e de difícil interpretação.

3.4.1 Hipóteses Testadas nos Modelos com Três Fatores Hierarquizados

Nesta seção, formalizam-se e discutem-se as hipóteses estatísticas sobre o fator A, fator B aninhado sob o fator A e fator C aninhado sob o fator B(A) que são usualmente testadas pelo procedimento GLM do SAS.

3.4.1.1 Hipóteses sobre o fator A

Embora, as somas de quadrados obtidas através da redução nas somas de quadrados, $R(\mu_i | \mu)$ e $R(\alpha | \mu)$, 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, independente dos dados serem desbalanceados e da ocorrência ou não de caselas vazias. Nesse caso, testam a hipótese do tipo I sobre as médias ponderadas não ajustadas. Sua forma geral é dada por:

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

Em modelos com três fatores hierarquizados, a soma de quadrados do tipo II fornecida pelo SAS-GLM, não testa hipótese do tipo II sobre as médias ponderadas do fator A ajustadas para o fator B, como nos casos de esquemas fatoriais, pois o fator B está aninhado sob o fator A. Sendo assim, a hipótese associada à soma de quadrados do tipo II é equivalente à hipótese 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). Assim, ela testa a hipótese do tipo III sobre as médias não ponderadas, dada por:

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

Se, entretanto, o número de níveis do fator C(A B) são diferentes, independentemente dos dados serem balanceados ou não, a soma de quadrados do tipo III fornecida pelo SAS-GLM 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:

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

De acordo com Searle (1987), as somas de quadrados dos tipos I, II e III são derivadas 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 a hipótese $H_0^{(2)}$.

3.4.1.2 Hipóteses sobre o fator B aninhado sob o fator A

No modelo com três fatores hierarquizados, as hipóteses testadas pelas somas de quadrados, $R(\mu_{ij} | \mu_i)$, $R[\beta(\alpha) | \mu, \alpha]$ e $SQB(A)$, correspondem às hipóteses

testadas pelas somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM, pois o fator B está aninhado sob o fator A. Nesse caso, independentemente dos dados serem desbalanceados ou não e do fator C(A B) ter ou não número de níveis diferentes, elas testam a hipótese do tipo I sobre as médias ponderadas não ajustadas, dada por:

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

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:

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

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

As hipóteses testadas através de, $R(\mu_{ijk} | \mu_{ij})$, $R[\gamma(\alpha | \beta) | \mu, \alpha, \beta(\alpha)]$ e $SQC(A B)$ correspondem 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 geral é dada por:

$$H_{oj}^{(6)} : \left\{ \left(\mu_{ijk} = \mu_{ijk'}, \forall k, k' (k \neq k') \right) \text{ dentro de } j \right\} \text{ dentro de } i.$$

3.4.2 Hipóteses Testadas nos Modelos com Estrutura do Tipo

Fatorial-Hierárquica

Aqui, são formalizadas e discutidas as hipóteses estatísticas sobre o fator A, fator B, fator C aninhado sob o fator B e das interações A*B e A*C(B), usualmente testadas pelo procedimento GLM do SAS nos modelos com estrutura do tipo fatorial-hierárquica.

3.4.2.1 Hipóteses sobre o fator A

Quando os dados são balanceados e se o fator C(B) tem o mesmo número de níveis em cada nível do fator B, então, tanto as hipóteses testadas pelas somas de quadrados, $R(\mu_i | \mu)$, $R(\alpha | \mu)$ e SQA, como as hipóteses testadas pelas somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM são equivalentes. Testam, ademais, hipóteses usuais de fácil interpretação.

Porém, se os dados são desbalanceados em presença ou não de caselas vazias, independentemente dos níveis do fator C(B) serem diferentes ou não, os usuários do SAS-GLM devem estar atentos quanto à ordem de entrada dos fatores no modelo, pois o SAS-GLM pode testar hipóteses distintas para a soma de quadrados do tipo I. Isto ocorre porque a soma de quadrados do tipo I fornecida pelo SAS-GLM é ajustada de acordo com a ordenação *a priori*, ou seja, cada efeito é ajustado para os efeitos que o precedem, ignorando os demais efeitos que ainda não entraram no modelo.

Assim, por exemplo, se for considerada a ordem [A, B, A*B, C(B) e A*C(B)] e se os dados são desbalanceados com todas as caselas ocupadas, as somas de quadrados obtidas através de, SQA, $R(\mu_i | \mu)$ e $R(\alpha | \mu)$, são equivalentes à soma de quadrados do tipo I fornecida pelo SAS-GLM e, portanto, testam a hipótese do tipo I sobre as médias ponderadas não ajustadas, dada por:

$$H_o^{(7)} : \sum_{j=1}^b \sum_{k=1}^{c_j} \frac{n_{ijk} \mu_{ijk}}{n_{i..}} = \sum_{j=1}^b \sum_{k=1}^{c_j} \frac{n_{i'jk} \mu_{i'jk}}{n_{i'..}}, \forall i, i' (i \neq i').$$

Se existem caselas vazias, elas ainda testam a hipótese $H_o^{(7)}$, mas consideram apenas as médias das caselas que estão ocupadas.

Agora, se for considerada a ordem [B, A, A*B, C(B) e A*C(B)], então, a soma de quadrados do tipo I, $R(\alpha | \mu, \beta)$, fornecida pelo SAS-GLM não testa a hipótese do tipo I e nem do tipo II, mas uma hipótese sobre as médias ponderadas ajustadas para o fator B. Nesse caso, se o fator A tem apenas dois níveis, sua forma geral é dada por:

$$H_o^{(8)} : \sum_{j=1}^b n_{ij.} \sum_{i' \neq i}^a n_{i'j.} (\tilde{\mu}_{ij.} - \tilde{\mu}_{i'j.}) / n_{.j.} = 0, \forall i, i', \text{ onde } \tilde{\mu}_{ij.} = \sum_{k=1}^{c_j} \frac{n_{ijk} \mu_{ijk}}{n_{ij.}}$$

No caso de caselas vazias, a soma de quadrados do tipo I fornecida pelo SAS-GLM testa a hipótese $H_o^{(8)}$, considerando apenas as médias das caselas que estão ocupadas.

Se, entretanto, o fator A tem mais de dois níveis, a soma de quadrados do tipo I testa uma hipótese não usual de difícil interpretação, baseada em funções estimáveis complexas.

A soma de quadrados do tipo II independe da ordem de entrada dos fatores no modelo, pois ela é ajustada para todos os fatores e interações, exceto interações que envolvem o fator de interesse e exceto fatores aninhados sob o fator de interesse. Assim, no modelo com a estrutura do tipo fatorial-hierárquica, se os dados são desbalanceados com todas as caselas ocupadas, a soma de quadrados do tipo II, $R[\alpha | \mu, \beta, \gamma(\beta)]$, fornecida pelo SAS-GLM, testa a hipótese do tipo II sobre as médias ponderadas ajustadas para os fatores B e C(B). Sua forma geral para o caso do fator A ter apenas dois níveis é:

$$H_o^{(9)} : \sum_{j=1}^b \sum_{k=1}^{c_j} n_{ijk} \sum_{i' \neq i}^a n_{i'jk} (\mu_{ijk} - \mu_{i'jk}) / n_{\bullet jk} = 0, \forall i, i'.$$

Se existem caselas vazias, então, a soma de quadrados do tipo II ainda testa a hipótese $H_o^{(9)}$, mas considera apenas as médias de caselas das linhas que estão completas.

Porém, se o fator A tem mais de dois níveis, a soma de quadrados do tipo II testa uma hipótese não usual de difícil interpretação baseada em funções estimáveis complexas do tipo II, independentemente de ocorrer caselas vazias ou não.

Quando os dados são desbalanceados com todas as caselas ocupadas e o fator C(B) tem o mesmo número de níveis dentro de cada nível do fator B, então, 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:

$$H_o^{(10)} : \mu_{i\bullet\bullet} = \mu_{i'\bullet\bullet}, \forall i, i' (i \neq i').$$

Se, no entanto, os níveis do fator C(B) são diferentes, quer os dados sejam ou não balanceados, a soma de quadrados do tipo III fornecida pelo SAS-GLM não testa a hipótese $H_o^{(10)}$, mas uma hipótese não usual, gerada a partir de funções estimáveis complexas do tipo III, dada por:

$$H_o^{(11)} : \frac{\sum_{j=1}^b \sum_{k=1}^{c_j} [\mu_{ijk} / (c_j + 1)]}{\sum_{j=1}^b [c_j / (c_j + 1)]} = \frac{\sum_{j=1}^b \sum_{k=1}^{c_j} [\mu_{i'jk} / (c_j + 1)]}{\sum_{j=1}^b [c_j / (c_j + 1)]}, \forall i, i' (i \neq i').$$

Quando existem caselas vazias, e se o fator A tem apenas dois níveis, então, a soma de quadrados do tipo III fornecida pelo SAS-GLM ainda testa a hipótese $H_o^{(11)}$, considerando apenas as médias de caselas das linhas que estão completas. Porém, se o fator A tem mais de dois níveis, a soma de quadrados do tipo III testa hipótese não usual, baseada em funções estimáveis complexas do tipo III de difícil interpretação.

Já a soma de quadrados do tipo IV testa hipótese $H_o^{(10)}$ se todas as caselas estão ocupadas, independentemente dos níveis do fator C(B) serem diferentes ou não. Se existem caselas vazias, como a soma de quadrados do tipo IV é gerada pela própria rotina do SAS-GLM, com base nas caselas ocupadas, portanto, pode testar tanto a hipótese $H_o^{(10)}$, considerando apenas para as médias das caselas que estão ocupadas, como pode testar outras hipóteses do tipo IV, eliminando caselas ocupadas nas linhas onde ocorreram as caselas vazias.

3.4.2.2 Hipóteses sobre o fator B

Se os dados são balanceados, independentemente dos níveis do fator C(B) serem diferentes ou não, as hipóteses testadas pelas somas de quadrados obtidas através de, SQB, $R(\mu_i + \beta_j \mid \mu_i)$ e $R(\beta \mid \mu, \alpha)$, correspondem às hipóteses testadas pelas somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM e, portanto, testam hipóteses usuais de fácil interpretação.

Se, entretanto, os dados são desbalanceados em presença ou não de caselas vazias, e se for considerada a ordem [A, B, A*B, C(B) e A*C(B)], então, as somas de quadrados dos tipos I e II, $R(\beta \mid \mu, \alpha)$, fornecidas pelo SAS-GLM são equivalentes e testam a hipótese do tipo II sobre as médias ponderadas ajustadas somente para o fator A, pois o fator C(B) está aninhado sob o fator B. Sua forma geral, no caso do fator B ter apenas dois níveis, é dada por:

$$H_o^{(12)} : \sum_{i=1}^a n_{ij\cdot} \sum_{j' \neq j}^b n_{ij'\cdot} (\tilde{\mu}_{ij\cdot} - \tilde{\mu}_{ij'\cdot}) / n_{i\cdot\cdot} = 0, \forall j, j', \text{ onde } \tilde{\mu}_{ij\cdot} = \sum_{k=1}^{c_j} \frac{n_{ijk} \mu_{ijk}}{n_{ij\cdot}}$$

Agora, se o fator B tem mais de dois níveis, as somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM testam hipóteses equivalentes, mas de difícil interpretação, geradas a partir de funções estimáveis complexas do tipo II.

Considerando a ordem [B, A, A*B, C(B) e A*C(B)], a soma de quadrados do tipo II ainda testa a hipótese $H_o^{(12)}$, porém a soma de quadrados do tipo I, $R(\beta \mid \mu)$,

fornecida pelo SAS-GLM testa hipótese do tipo I sobre as médias ponderadas não ajustadas, equivalente àquela obtida através do procedimento usual, SQB, dada por:

$$H_o^{(13)} : \sum_{i=1}^a \sum_{k=1}^{c_j} \frac{n_{ijk} \mu_{ijk}}{n_{\bullet j \bullet}} = \sum_{i=1}^a \sum_{k=1}^{c_{j'}} \frac{n_{ij'k} \mu_{ij'k}}{n_{\bullet j' \bullet}}, \forall j, j' (j \neq j').$$

Se os dados são desbalanceados com todas as caselas ocupadas, quer os níveis do fator C(B) sejam diferentes ou não, as somas de quadrados dos tipos III e IV são equivalentes àquela obtida através do método dos quadrados de médias ponderadas de Yates (1934) e, portanto, testam a hipótese do tipo III sobre as médias não ponderadas. Sua forma geral é:

$$H_o^{(14)} : \mu_{\bullet j \bullet} = \mu_{\bullet j' \bullet}, \forall j, j' (j \neq j').$$

No caso de caselas vazias, as somas de quadrados dos tipos III e IV testam hipóteses não usuais de difícil interpretação, geradas a partir de funções estimáveis complexas dos tipos III e IV.

3.4.2.3 Hipóteses sobre a Interação A*B

Para a interação A*B, se os dados são balanceados, independentemente dos níveis do fator C(B) serem diferentes ou não, tanto as hipóteses testadas pelas somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM, como as hipóteses

testadas pelas somas de quadrados, $R(\mu_{ij} | \mu_i + \beta_j)$, $R[(\alpha\beta) | \mu, \alpha, \beta]$ e $SQ(A*B)$, são equivalentes e de fácil interpretação.

Se, no entanto, os dados são desbalanceados com todas as caselas ocupadas, tanto as somas de quadrados, $R(\mu_{ij} | \mu_i + \beta_j)$ e $R[(\alpha\beta) | \mu, \alpha, \beta]$, como a soma de quadrados do tipo I fornecida pelo SAS-GLM, testa hipóteses equivalentes sobre a interação entre as médias ponderadas não ajustadas, dada por:

$$H_o^{(15)} : \sum_{k=1}^{c_j} \frac{n_{ijk} \mu_{ijk}}{n_{ij\bullet}} - \sum_{k=1}^{c_{j'}} \frac{n_{ij'k} \mu_{ij'k}}{n_{ij'\bullet}} - \sum_{k=1}^{c_j} \frac{n_{i'jk} \mu_{i'jk}}{n_{i'j\bullet}} + \sum_{k=1}^{c_{j'}} \frac{n_{i'j'k} \mu_{i'j'k}}{n_{i'j'\bullet}} = 0,$$

$$\forall i, i', j, j' (i \neq i', j \neq j').$$

Já a hipótese testada pela soma de quadrados obtida através do procedimento usual, $SQ(A*B)$, não é definida e, de acordo com Searle (1971), não corresponde à hipótese usual sobre a interação.

Segundo Elston & Bush (1964), Searle (1987), Iemma (1993) entre outros, quando existem caselas vazias, as hipóteses sobre as interações devem ser interpretadas com cautela, pois elas testam que apenas algumas bem determinadas interações são nulas. Assim, no caso de caselas vazias, tanto as somas de quadrados, $R(\mu_{ij} | \mu_i + \beta_j)$ e $R[(\alpha\beta) | \mu, \alpha, \beta]$, como a soma de quadrados do tipo I fornecida pelo SAS-GLM ainda testam uma hipótese semelhante à $H_o^{(15)}$, mas consideram apenas os componentes da interação, relativos às caselas que estão ocupadas.

Se os dados são desbalanceados em presença ou não de caselas vazias, a soma de quadrados do tipo II testa hipótese não usual de difícil interpretação, gerada a partir de funções estimáveis complexas do tipo II.

As somas de quadrados dos tipos III e IV testam hipóteses equivalentes sobre a interação entre as médias não ponderadas, se os dados são desbalanceados com todas as caselas ocupadas. Sua forma geral é dada por:

$$H_o^{(16)} : \mu_{ij\bullet} - \mu_{ij'\bullet} - \mu_{i'j\bullet} + \mu_{i'j'\bullet} = 0, \forall i, i', j, j' (i \neq i', j \neq j').$$

Se, no entanto, existem caselas vazias e se o fator A tem apenas dois níveis, então, as somas de quadrados dos tipos III e IV testam a hipótese $H_o^{(16)}$, mas consideram apenas as médias de caselas das linhas que estão completas.

Agora, se o fator A tem mais de dois níveis, então, a soma de quadrados do tipo III testa hipótese não usual de difícil interpretação, baseada em funções estimáveis complexas do tipo III. A soma de quadrados do tipo IV pode testar tanto a hipótese $H_o^{(16)}$, apenas para as médias de caselas das linhas que estão completas, como pode testar outras hipóteses complexas do tipo IV.

3.4.2.4 - Hipóteses sobre o fator C aninhado sob o fator B

Quando os dados são balanceados, independente de ocorrer níveis diferentes para o fator C(B), tanto as somas de quadrados dos tipos I, II, III e IV fornecidas

pelo SAS-GLM, como as somas de quadrados, $R(\mu_{ij} + \gamma_{k(j)} | \mu_{ij})$, $R[\gamma | \mu, \alpha, \beta, (\alpha\beta)]$ e $SQC(B)$, testam hipóteses equivalentes de fácil interpretação.

Se, no entanto, os dados são desbalanceados com todas as caselas ocupadas e se o fator C(B) tem apenas dois níveis, então, tanto as somas de quadrados dos tipos I e II, como as somas de quadrados, $R(\mu_{ij} + \gamma_{k(j)} | \mu_{ij})$, $R[\gamma | \mu, \alpha, \beta, (\alpha\beta)]$ testam hipóteses equivalentes da forma:

$$H_o^{(17)}: \left\{ \sum_{i=1}^a \sum_{j=1}^b n_{ijk} \sum_{k' \neq k}^{c_j} n_{ijk'} (\mu_{ijk} - \mu_{ijk'}) / (n_{ijk} + n_{ijk'}) = 0, \forall k, k' \right\} \text{ dentro de cada } j.$$

Se existem caselas vazias, as somas de quadrados dos tipos I e II testam tanto a hipótese $H_o^{(17)}$, apenas para as médias de caselas das colunas completas do fator C(B), como podem testar outras hipóteses não usuais, baseadas em funções estimáveis complexas.

Quando os níveis do fator C(B) são maiores que dois, as somas de quadrados dos tipos I e II testam hipóteses equivalentes, mas não usuais de difícil interpretação, baseadas em funções estimáveis complexas, independentemente de ocorrer ou não caselas vazias.

No caso de dados desbalanceados com todas as caselas ocupadas, as somas de quadrados dos tipos III e IV são equivalentes ao método dos quadrados de médias ponderadas de Yates (1934) e testam a hipótese do tipo III sobre as médias não ponderadas, dada por:

$$H_o^{(18)}: \left\{ \mu_{\bullet jk} = \mu_{\bullet jk'} , \quad \forall k, k' (k \neq k') \right\} \text{ dentro de } j .$$

Se existem caselas vazias, as somas de quadrados dos tipos III e IV testam tanto a hipótese $H_o^{(18)}$, apenas para as médias de caselas das colunas completas do fator C(B), como podem testar outras hipóteses não usuais, baseadas em funções estimáveis complexas.

Já a soma de quadrados obtida através do procedimento usual, SQC(B), testa a hipótese do tipo I sobre as médias ponderadas não ajustadas,

$$H_o^{(19)} : \left\{ \sum_{i=1}^a \frac{n_{ijk} \mu_{ijk}}{n_{\bullet jk}} = \sum_{i=1}^a \frac{n_{ijk'} \mu_{ijk'}}{n_{\bullet jk'}}, \quad \forall k, k' (k \neq k') \right\} \text{ dentro de } j .$$

3.4.2.5 Hipótese sobre a interação A*C(B)

Em modelos com estrutura do tipo fatorial-hierárquica, se os dados são desbalanceados com todas as caselas ocupadas, independentemente dos níveis do fator C(B) serem diferentes ou não, tanto as somas de quadrados, $R(\mu_{ijk} \mid \mu_{ij} + \gamma_{k(j)})$ e $R[(\alpha\gamma) \mid \mu, \alpha, \beta, (\alpha\beta), \gamma]$, como as somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM, testam hipóteses equivalentes da forma:

$$H_{oj}^{(20)}: \left\{ \mu_{ijk} - \mu_{ijk'} - \mu_{i'jk} + \mu_{i'jk'} = 0 , \quad \forall i, i', k, k' (i \neq i', k \neq k') \right\} \text{ dentro de } j .$$

Se existem caselas vazias, elas ainda testam a hipótese $H_o^{(20)}$, mas consideram apenas os componentes da interação que envolvem todas as caselas ocupadas.

Já a hipótese testada pela soma de quadrados obtida através do procedimento usual, $SQA * C(B)$, também não é definida e, portanto, não corresponde à hipótese usual sobre a interação $A * C(B)$.

4 RESULTADOS E DISCUSSÃO

Com base nos conjuntos de dados dos exemplos 1 e 2, discutem-se os resultados fornecidos pelo SAS-GLM visando interpretar as hipóteses estatísticas e as somas de quadrados a elas associadas, nos modelos com as estruturas do tipo estritamente hierárquica e do tipo fatorial-hierárquica. Também são comparadas, com base nos resultados fornecidos por outros *softwares* estatísticos muito utilizados pelos pesquisadores das ciências aplicadas, as hipóteses testadas por cada um deles.

4.1 Modelos com Três Fatores Hierarquizados

Exemplo 1: Conjunto de dados sobre comprimento de fibras de eucalipto, adaptado de Padovani (1984) e reproduzido na Tabela 3. 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 seções 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, 3º e 2º anos de crescimento. Para cada árvore, foram obtidas três amostras correspondentes às posições 1, 2 e 3, relativas aos anos de crescimento. A partir do 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 desbalançamento com caselas vazias. Sendo assim, foram consideradas apenas duas posições na árvore 2 e número diferente de lâminas.

Tabela 3 - Comprimento de fibras de eucalipto, em micra.

Árvore										
A ₁						A ₂				
Posição						Posição				
B ₁		B ₂	B ₃			B ₁			B ₂	
Lâm.		Lâm.	Lâm.			Lâm.			Lâm.	
C ₁	C ₂	C ₁	C ₁	C ₂	C ₃	C ₁	C ₂	C ₃	C ₁	C ₂
0,791	0,649	1,729	1,162	0,909	0,979	0,944	0,702	0,708	0,968	1,210
0,749	0,791	0,915	0,785	1,062	0,962	0,820	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).

4.1.1 Obtenção das Somas de Quadrados

Os modelos definidos em (3) de 3.2.1 e (8) de 3.2.2, assumem diversas caracterizações conforme o interesse. Neste caso, com base no conjunto de dados da

Tabela 3 e procedendo-se as parametrizações sucessivas e ordenadas, como em (M.11) até (M.14) ou como em (S.11) até (S.14) de 3.2.3, foram obtidas as somas de quadrados sequenciais, apresentadas a seguir.

$$R(\mu) = R(\mu) = 28,7444895$$

$$R(\mu_i | \mu) = R(\alpha | \mu) = 0,1016466$$

$$R(\mu_{ij} | \mu_i) = R[\beta(\alpha) | \mu, \alpha] = 0,5514341$$

$$R(\mu_{ijk} | \mu_{ij}) = R[\gamma(\alpha \beta) | \mu, \alpha, \beta(\alpha)] = 0,0331818$$

Para realizar a análise de variância através do procedimento GLM do sistema estatístico SAS, utilizou-se o programa 1:

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 1,210
2 2 2 0,915
;
PROC GLM;
CLASS A B C;
MODEL Y = A B(A) C(A B) / SS1 SS2 SS3 SS4 E1 E2 E3 E4;
RUN;
```

Os resultados constam da Tabela 4, na qual pode ser observada a equivalência entre as somas de quadrados dos tipos I e II, com as somas de quadrados obtidas por parametrização sequencial do modelo.

Tabela 4 - 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	<u>0,02167112</u>	0,03163605	
B(A)	3	0,55143413	0,52756684	0,52756684	
C(A B)	6	0,03318183	0,03318183	0,03318183	

4.1.2 Hipóteses Testadas pelo SAS-GLM nos Modelos com Três Fatores Hierarquizados

4.1.2.1 Hipóteses sobre o fator A

As somas de quadrados referentes ao fator A, obtidas através da redução nas somas de quadrados, $R(\mu_i | \mu)$ e $R(\alpha | \mu)$, bem como aquela obtida através do procedimento usual, SQA, não especificam as hipóteses testadas. Porém, elas correspon-

dem às hipóteses testadas pelas somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Assim, para os dados da Tabela 3, elas testam a hipótese do tipo I sobre as médias ponderadas não ajustadas, dada em 3.4.1.1, para o modelo de médias de caselas, por:

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

A hipótese $H_o^{(1)}$ pode ser descrita na forma $H_o^{(1)} : B'\mu = \phi$ 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 $H_o^{(1)}$ pode ser obtida, utilizando a estatística de Wald, $SQH_o^{(1)} = (B'\hat{\mu})'[B'(W'W)^{-1}B]^{-1}(B'\hat{\mu})$ onde, das equações normais dadas em 3.2.1,

$$\hat{\mu}' = [0,77 \ 0,7337 \ 1,1257 \ 0,9735 \ 0,9972 \ 1,0532 \ 0,8357 \ 0,7573 \ 0,752 \ 0,9857 \ 1,0625];$$

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

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

Desse modo,

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

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

$$H_o^{(1)} : \left\{ \alpha_1 + \frac{1}{19} (5\beta_{1(1)} + 4\beta_{2(1)} + 10\beta_{3(1)} + 2\gamma_{1(11)} + 3\gamma_{2(11)} + 4\gamma_{1(12)} + 2\gamma_{1(13)} + 4\gamma_{2(13)} + 4\gamma_{3(13)}) \right\} =$$

$$= \left\{ \alpha_2 + \frac{1}{15} (10\beta_{1(2)} + 5\beta_{2(2)} + 3\gamma_{1(21)} + 3\gamma_{2(21)} + 4\mu_{3(21)} + 3\gamma_{1(22)} + 2\gamma_{2(22)}) \right\},$$

Tabela 5 - Funções estimáveis para o fator A fornecidas pelo SAS-GLM.

General Linear Models Procedure				
Estimable Functions for: A				
Effects		Type I e II	Type III	Type IV
		Coefficients	Coefficients	Coefficients
Intercept		0	0	0
A	1	L2	L2	L2
	2	-L2	-L2	-L2
B(A)	11	0,2632*L2	0,3478*L2	0,3333*L2
	21	0,2105*L2	0,2609*L2	0,1667*L2
	31	0,5263*L2	0,3913*L2	0,5*L2
	12	-0,6667*L2	-0,5294*L2	-0,6*L2
	22	-0,3333*L2	-0,4706*L2	-0,4*L2
C(A B)	111	0,1053*L2	0,1739*L2	0,1667*L2
	211	0,1579*L2	0,1739*L2	0,1667*L2
	112	0,2105*L2	0,2609*L2	0,1667*L2
	113	0,1053*L2	0,1304*L2	0,1667*L2
	213	0,2105*L2	0,1304*L2	0,1667*L2
	313	0,2105*L2	0,1304*L2	0,1667*L2
	121	-0,2*L2	-0,1765*L2	-0,2*L2
	221	-0,2*L2	-0,1765*L2	-0,2*L2
	321	-0,2667*L2	-0,1765*L2	-0,2*L2
	122	-0,2*L2	-0,2353*L2	-0,2*L2
	222	-0,1333*L2	-0,2353*L2	-0,2*L2

como pode ser observado na Tabela 5, fornecida pelo SAS-GLM, fazendo-se $L2 = 1$, onde cada L está associada ao número de graus de liberdade (g.l.) do fator em estudo. Sendo assim, a Tabela 5 referente as funções estimáveis dos tipos I, II, III e IV, consta somente de $L2$, pois o fator A tem 1 g.l. Sem dúvida, é uma hipótese difícil de ser interpretada, mesmo com apenas 1 g.l., especialmente aos pesquisadores não iniciados na teoria dos testes de hipóteses estatísticas.

Como o fator C tem 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, portanto, não é equivalente à hipótese $H_o^{(2)}$. Ao contrário, ela testa hipótese não usual, baseada na soma de quadrados gerada a partir de funções estimáveis complexas do tipo III. Assim, para os dados da Tabela 3, tem-se conforme 3.4.1.1,

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

Portanto,

$$H_o^{(3)} : \frac{4(\mu_{111} + \mu_{112}) + 6\mu_{121} + 3(\mu_{131} + \mu_{132} + \mu_{133})}{23} = \frac{3(\mu_{211} + \mu_{212} + \mu_{213}) + 4(\mu_{221} + \mu_{222})}{17} .$$

Através da estatística de Wald, tem-se:

$$B' = [4/23 \ 4/23 \ 6/23 \ 3/23 \ 3/23 \ 3/23 \ -3/17 \ -3/17 \ -3/17 \ -4/17 \ -4/17] ;$$

$$B'\hat{\mu} = 0,0538708 ; [B'(W'W)^{-1}B] = 0,1339141 \text{ e } [B'(W'W)^{-1}B]^{-1} = 7,4674754.$$

Logo,

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

Em termos de modelo superparametrizado, a hipótese $H_o^{(3)}$ resulta em,

$$H_o^{(3)} : \left\{ \alpha_1 + \frac{1}{23} (8\beta_{1(1)} + 6\beta_{2(1)} + 9\beta_{3(1)} + 4\gamma_{1(11)} + 4\gamma_{2(11)} + 6\gamma_{1(12)} + \right. \\ \left. + 3\gamma_{1(13)} + 3\gamma_{2(13)} + 3\gamma_{3(13)}) \right\} = \\ = \left\{ \alpha_2 + \frac{1}{17} (9\beta_{1(2)} + 8\beta_{2(2)} + 3\gamma_{1(21)} + 3\gamma_{2(21)} + 3\gamma_{3(21)} + 4\gamma_{1(22)} + 4\gamma_{2(22)}) \right\} ,$$

como se observa na Tabela 5, fazendo-se $L2 = 1$, também associada a 1 g.l..

Realmente, como comentado anteriormente, o modelo superparametrizado satisfaz os anseios dos pesquisadores, exibindo explicitamente os parâmetros de interesse, porém suas hipóteses são extremamente difíceis de interpretar. Além disso, a soma de quadrados do tipo III fornecida pelo SAS-GLM não testa a hipótese sobre as médias não ponderadas, $H_o^{(2)}$, mas uma hipótese completamente diferente, baseada em funções estimáveis complexas do tipo III. Embora a hipótese testada pela soma de quadrados do tipo III tenha sido formulada neste trabalho, mesmo assim, este resultado vem confirmar a preocupação de Searle (1994) de que, parece não haver uma explicação estatística para

que a soma de quadrados do tipo III fornecida pelo SAS-GLM seja calculada dessa forma, tanto nos manuais de utilização como na literatura estatística.

Segundo Hocking (1985), a ocorrência de soma de quadrados do tipo III que não testa a hipótese $H_o^{(2)}$, mas uma hipótese não usual derivada de funções estimáveis complexas do número de observações nas caselas, é consequência do procedimento computacional utilizado pelo SAS-GLM, sem outra explicação. Para o autor, isso ocorre freqüentemente em modelos que envolvem três ou mais fatores de efeitos fixos.

Nos modelos com três fatores hierarquizados, a soma de quadrados do tipo IV testa sempre a hipótese $H_o^{(2)}$ sobre as médias não ponderadas, independentemente de ocorrerem caselas vazias ou não. Assim, tem-se:

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

e então,

$$B' = [1/6 \ 1/6 \ 1/6 \ 1/6 \ 1/6 \ 1/6 \ -1/5 \ -1/5 \ -1/5 \ -1/5 \ -1/5] ;$$

$$B'\hat{\mu} = 0,0636028 ; [B'(W'W)^{-1}B] = 0,1278704 \text{ e } [B'(W'W)^{-1}B]^{-1} = 7,82042.$$

Sendo assim,

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

Procedendo-se de modo análogo ao anterior, obtém-se a forma de $H_o^{(2)}$ no modelo superparametrizado,

$$H_o^{(2)} : \left\{ \alpha_1 + \frac{1}{6}(2\beta_{1(1)} + \beta_{2(1)} + 3\beta_{3(1)} + \gamma_{1(11)} + \gamma_{2(11)} + \gamma_{1(12)} + \right. \\ \left. + \gamma_{1(13)} + \gamma_{2(13)} + \gamma_{3(13)}) \right\} = \\ = \left\{ \alpha_2 + \frac{1}{5}(3\beta_{1(2)} + 2\beta_{2(2)} + \gamma_{1(21)} + \gamma_{2(21)} + \gamma_{3(21)} + \gamma_{1(22)} + \gamma_{2(22)}) \right\},$$

como se observa na Tabela 5, fazendo-se $L2 = 1$.

4.1.2.2 Hipóteses sobre o fator B aninhado sob o fator A

As hipóteses sobre o fator B, agora associada a 3 g.l., testadas pelas somas de quadrados, $R(\mu_{ij} | \mu_i)$, $R[\beta(\alpha) | \mu, \alpha]$ e $SQB(A)$, são equivalentes às hipóteses testadas pelas somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Assim, elas testam a hipótese do tipo I. Conforme descrito em 3.4.1.2, uma forma para $H_o^{(4)}$ é:

$$H_o^{(4)} : \left\{ \begin{array}{l} \frac{(2\mu_{111} + 3\mu_{112})}{5} = \frac{(2\mu_{131} + 4\mu_{132} + 4\mu_{133})}{10} \\ \mu_{121} = \frac{(2\mu_{131} + 4\mu_{132} + 4\mu_{133})}{10} \\ \frac{(3\mu_{211} + 3\mu_{212} + 4\mu_{213})}{10} = \frac{(3\mu_{221} + 2\mu_{222})}{5} \end{array} \right.$$

Desse modo, a soma de quadrados associada à hipótese $H_o^{(4)}$, pode ser

obtida através da estatística de Wald onde,

$$B'_1 = [2/5 \ 3/5 \ 0 \ -2/10 \ -4/10 \ -4/10 \ 0 \ 0 \ 0 \ 0 \ 0] ;$$

$$B'_2 = [0 \ 0 \ 1 \ -2/10 \ -4/10 \ -4/10 \ 0 \ 0 \ 0 \ 0 \ 0] ;$$

$$B'_3 = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 3/10 \ 3/10 \ 4/10 \ -3/5 \ -2/5] ;$$

$$B' = [B'_1 | B'_2 | B'_3]' ; B'\hat{\mu} = [-0,2667 \ 0,11085 \ -0,2377]' ;$$

$$[B'(W'W)^{-1}B] = \begin{bmatrix} 3/10 & 1/10 & 0 \\ & 7/20 & 0 \\ (\text{Sim.}) & & 3/10 \end{bmatrix}$$

$$\text{e } [B'(W'W)^{-1}B]^{-1} = \begin{bmatrix} 70/19 & -20/19 & 0 \\ & 60/19 & 0 \\ (\text{Sim.}) & & 10/3 \end{bmatrix}.$$

Logo,

$$SQH_o^{(4)} = 0,5514341 = R [\beta(\alpha) | \mu, \alpha].$$

Descrevendo em termos de modelo superparametrizado, a hipótese $H_o^{(4)}$ resulta em,

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

como se observa na Tabela 6, 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..

Embora não existam os termos α_i 's nas expressões da hipótese $H_0^{(4)}$, pois eles foram ajustados, verifica-se ainda que elas são de difícil interpretação.

Tabela 6 - Funções estimáveis para os fatores B(A) e C(A B) fornecidas pelo SAS-GLM.

General Linear Models Procedure				
Estimable Functions for:				
		B(A)		C(A B)
		Type I e II	Type III e IV	Type I, II, III e IV
Effects		Coefficients	Coefficients	Coefficients
Intercept		0	0	0
A	1	0	0	0
	2	0	0	0
B(A)	11	L4	L4	0
	21	L5	L5	0
	31	-L4-L5	-L4-L5	0
	12	L7	L7	0
	22	-L7	-L7	0
C(A B)	111	0,4*L4	0,5*L4	L9
	211	0,6*L4	0,5*L4	-L9
	112	L5	L5	0
	113	-0,2*L4-0,2*L5	-0,3333*L4-0,3333*L5	L12
	213	-0,4*L4-0,4*L5	-0,3333*L4-0,3333*L5	L13
	313	-0,4*L4-0,4*L5	-0,3333*L4-0,3333*L5	-L12-L13
	121	0,3*L7	0,3333*L7	L15
	221	0,3*L7	0,3333*L7	L16
	321	0,4*L7	0,3333*L7	-L15-L16
	122	-0,6*L7	-0,5*L7	L18
	222	-0,4*L7	-0,5*L7	-L18

Já as somas de quadrados dos tipos III e IV fornecidas pelo SAS-GLM testam a hipótese do tipo III sobre as médias não ponderadas. Uma forma para a hipótese $H_0^{(5)}$ é:

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

então,

$$B'_1 = [1/2 \ 1/2 \ 0 \ -1/3 \ -1/3 \ -1/3 \ 0 \ 0 \ 0 \ 0];$$

$$B'_2 = [0 \ 0 \ 1 \ -1/3 \ -1/3 \ -1/3 \ 0 \ 0 \ 0 \ 0];$$

$$B'_3 = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/3 \ 1/3 \ 1/3 \ -1/2 \ -1/2];$$

$$B' = [B'_1 | B'_2 | B'_3]'; \quad B'\hat{\mu} = [-0,256167 \ 0,11775 \ -0,242417]';$$

$$[B'(W'W)^{-1}B] = \begin{bmatrix} 0,319444 & 0,111111 & 0 \\ & 0,361111 & 0 \\ (\text{Sim.}) & & 0,310185 \end{bmatrix}$$

$$\text{e } [B'(W'W)^{-1}B]^{-1} = \begin{bmatrix} 3,505618 & -1,078652 & 0 \\ & 3,1011236 & 0 \\ (\text{Sim.}) & & 3,2238806 \end{bmatrix}.$$

Sendo assim,

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

Descrevendo a hipótese $H_o^{(5)}$ em termos de modelo superparametrizado, resulta em,

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

como se observa na Tabela 6, fazendo-se: L4=1 e L5=L7=0 ; L5=1 e L4=L7=0 ; L4=L5=0 e L7=1, pois está associada a 3 g.l..

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

Para o conjunto de dados da Tabela 3, as somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM, bem como aquelas obtidas através da notação- $R(\cdot)$, $R(\mu_{ijk} | \mu_{ij})$ e $R[\gamma(\alpha \beta) | \mu, \alpha, \beta(\alpha)]$, testam a hipótese $H_o^{(6)}$ descrita em 3.4.1.3.

Uma forma, agora associada a 6 g.l., é dada por:

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

como se observa na Tabela 6 fazendo-se: 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 e, finalmente, L18=1 e os demais iguais a zero. Aqui há 6 L's diferentes que representam 6 g.l. para o fator C(A B).

Desse modo, a soma de quadrados associada à hipótese $H_0^{(6)}$ pode ser obtida facilmente através da estatística de Wald onde,

$$B'_1 = [1 -1 0 0 0 0 0 0 0 0 0 0]; B'_4 = [0 0 0 0 0 0 1 0 -1 0 0];$$

$$B'_2 = [0 0 0 1 0 -1 0 0 0 0 0]; B'_5 = [0 0 0 0 0 0 0 1 -1 0 0];$$

$$B'_3 = [0 0 0 0 1 -1 0 0 0 0 0]; B'_6 = [0 0 0 0 0 0 0 0 0 1 -1];$$

$$B' = [B'_1 | B'_2 | B'_3 | B'_4 | B'_5 | B'_6]';$$

$$B'\hat{\mu} = [0,036333 \quad -0,07975 \quad -0,056 \quad 0,083667 \quad 0,005333 \quad -0,076833]';$$

$$[B'(W'W)^{-1}B] = \begin{bmatrix} 5/6 & 0 & 0 & 0 & 0 & 0 \\ & 3/4 & 1/4 & 0 & 0 & 0 \\ & & 1/2 & 0 & 0 & 0 \\ & & & 7/12 & 1/4 & 0 \\ & & & & 7/12 & 0 \\ \text{(Sim.)} & & & & & 5/6 \end{bmatrix}$$

$$\text{e } [B'(W'W)^{-1}B]^{-1} = \begin{bmatrix} 6/5 & 0 & 0 & 0 & 0 & 0 \\ & 8/5 & -4/5 & 0 & 0 & 0 \\ & & 12/5 & 0 & 0 & 0 \\ & & & 21/10 & -9/10 & 0 \\ & & & & 21/10 & 0 \\ \text{(Sim.)} & & & & & 6/5 \end{bmatrix}$$

Sendo assim,

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

4.1.3 Hipóteses Testadas por Outros *Softwares* nos Modelos com Três Fatores Hierarquizados

Conforme descrito anteriormente, com o objetivo de elucidar aos usuários, apresenta-se aqui uma comparação, sem o apelo de competição, as hipóteses testadas através do SAS-GLM com aquelas testadas por outros *Softwares* estatísticos. Foram utilizados o MINITAB, o NTIA, o STATGRAPHICS, o SAEG, o SANEST, o GLIM, o STATISTICA e o BMDP.

4.1.3.1 MINITAB

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, como pode ser observado nas Tabelas 4 e 7, quando os dados são desbalanceados com todas as caselas ocupadas, o procedimento GLM do MINITAB fornece somas de quadrados dos tipos sequenciais e ajustadas, equivalentes às somas de quadrados dos tipos I e III fornecidas pelo SAS-GLM. Entretanto, para o conjunto de dados da Tabela 3, o procedimento GLM fornece apenas somas de quadrados sequenciais 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, fato que o MINITAB interpreta como caselas vazias.

Tabela 7 - Análise de variância fornecida pelo MINITAB

General Linear Model			
Analysis of Variance for Y			
Source	Model DF	Reduced DF	Seq. SS
A	1	1	0,101647
B(A)	4	3+	0,551434
C(A B)	12	6+	0,033182
Error	16	23	0,780930
Total	33	33	1,467192

+ Rank deficiency due to empty cells, unbalanced nesting or collinearity. No storage of results or further analysis will be done.

4.1.3.2 NTIA

Programa 3: Programa NTIA para modelos com Três Fatores Hierarquizados

```
NTIA>GENESE NESTED
NTIA>NUM A B C Y;
NTIA>ARQUIVO M=ABREF(R:\ARVOHIE2.DAD) A B C Y;
NTIA>{LEIAF(M)};
NTIA>MODLIN NESTED
MOD Y = A B(A) 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 4). Testam, portanto, hipóteses equivalentes, exceto para a soma de quadrados do tipo parcial referente ao fator A (grifada na Tabela 8) 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, $R[\alpha^* | \mu^*, \beta^*, \gamma(\alpha \beta)^*] = R[\mu^*, \alpha^*, \beta^*, \gamma(\alpha \beta)^*] - R[\mu^*, \beta^*, \gamma(\alpha \beta)^*]$, obtida através do modelo reparametrizado propostos por Overall & Spiegel (1969), Carlson & Timm (1974) entre outros. De acordo com Searle *et al.* (1981), quando todas as caselas estão ocupadas ($n_{ijk} > 0$), a soma de quadrados $R[\alpha^* | \mu^*, \beta^*, \gamma(\alpha \beta)^*]$ 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 testes de hipóteses do tipo III.

Tabela 8 - Análise de variância fornecida pelo NTIA

F. de Variação	G.L	S.Q Seq.	Q.M	Valor F	PR > F
A	1	0,10164659	0,10164659	2,9937	0,097
B(A)	3	0,55143413	0,18381138	5,4136	0,006
C(A B)	6	0,03318183	0,00553031	0,1629	0,984
Resíduo	23	0,78092992	0,03395347		
Total	33	1,46719247			
F. de Variação	G.L	S.Q Parcial	Q.M	Valor F	PR > F
A	1	<u>0,02470823</u>	0,02470823	0,7277	0,402
B(A)	3	0,52756684	0,17585561	5,1793	0,007
C(A B)	6	0,03318183	0,00553031	0,1629	0,984

4.1.3.3 STATGRAPHICS

O procedimento NESTED do sistema STATGRAPHICS não realiza as análises para modelos hierarquizados com dados desbalanceados.

4.1.3.4 SAEG

O procedimento HIERÁRQUICO do SAEG realiza as análises para modelos hierarquizados, fornecendo somas de quadrados sequenciais 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.

Tabela 9 - Análise de Variância fornecida pelo SAEG

ANALISE DE VARIANCIA E COVARIANCIA					
K PARA COMPONENTES DE VARIANCIA E COVARIANCIA					
	C	B	A		
A	3,26	7,93	16,76		
B	3,22	6,08			
C	2,97				
F. VAR.	G.L	S.Q	Q.M	C. VAR.	%
TOTAL	33	1,4671920	0,0444604	0,0636585	100,00
A	1	0,1016466	0,1014660	-0,0081514	0,00
B	3	0,5514341	0,1838114	0,0297050	46,66
C	6	0,0331819	0,0055303	-0,0095808	0,00
RESIDUO	23	0,7809298	0,0339535	0,0339535	53,34

Embora o procedimento HIERÁRQUICO do SAEG seja indicado somente para modelos com estrutura do tipo estritamente hierárquica, não evidencia claramente aos usuários que os efeitos dos fatores B e C são hierarquizados.

4.1.3.5 SANEST

O SANEST não realiza as análises para modelos hierarquizados com dados desbalanceados.

4.1.3.6 GLIM

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
1 1 2 0,649
.
.
.
2 2 2 1,210
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 apenas somas de quadrados do tipo sequencial equivalentes às somas de quadrados do tipo I fornecidas pelo SAS-GLM, sem realizar nenhum teste estatístico. Conforme observa-se na Tabela 10, os resultados fornecidos

pelo GLIM, são extremamente difíceis de interpretar, pois não mostram claramente aos usuários, quais efeitos estão envolvidos e/ou o que está sendo testado.

Tabela 10 - Análise de variância fornecida pelo GLIM

[o]	GLIM 4, update 8 for IBM etc. 80386 PC / DOS on 12-Abr-1998
[o]	(copyright) 1992 Royal Statistical Society, London
[o]	
[o]	deviance = 1,4672
[o]	residual df = 33
[o]	
[o]	deviance = 1,3655 (change = -0,1016)
[o]	residual df = 32 (change = -1)
[o]	
[o]	deviance = 0,8141 (change = -0,5514)
[o]	residual df = 29 (change = -3)
[o]	
[o]	deviance = 0,7809 (change = -0,0332)
[o]	residual df = 23 (change = -6)

4.1.3.7 STATISTICA

O STATISTICA emite a mensagem “DESIGN INCOMPLETE ; TEST PLANNED COMPARISONS OR SPECIFIC EFFECTS” e não realiza as análises para modelos com três fatores hierarquizados com dados desbalanceados. 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, elas testam as hipóteses $H_0^{(2)}$, $H_0^{(5)}$ e $H_0^{(6)}$, conforme consta na Tabela 11.

Tabela 11 - Comparações planejados para os efeitos principais fornecidos pelo
STATISTICA

Planned Comparison - Fator A					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	0,031636	1	0,031636	0,931747	0,34445
Error	0,780930	23	0,033953		
Planned Comparison - Fator B(A)					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	0,527567	3	0,175856	5,179311	0,00700
Error	0,780930	23	0,033953		
Planned Comparison - Fator C(A B)					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	0,033182	6	0,005530	0,162879	0,984139
Error	0,780930	23	0,033953		

4.1.3.8 BMDP

Programa 5: Programa BMDP para modelos com Três Fatores Hierarquizados


```

/ INPUT   TITLE IS 'MODELO COM TRES FATORES HIERARQUIZADOS'.
          VARIABLES = 4.
          FORMAT = FREE.
/ VARIABLE NAMES = A, B, C, Y.
/ BETWEEN FACTORS = A, B, C.
          CODES(A) = 1,2.
          CODES(B) = 1 TO 5.
          CODES(C) = 1 TO 11.
/ WEIGHTS BETWEEN = EQUAL.
/ END
1 1 1 0,791
1 1 1 0,749
1 1 2 0,649
.
.
.
2 5 11 1,210
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 do SAS-GLM. Se existem caselas vazias, como no exemplo em questão, 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. Conforme observa-se na Tabela 12, o BMDP também não evidencia claramente

aos usuários que os efeitos envolvidos são hierarquizados.

Tabela 12 - Análise de variância fornecida pelo BMDP

BMDP4V--UNIVARIATE AND MULTIVARIATE ANALYSIS OF VA-
RIANCE AND COVARIANCE, INCLUDING REPEATED MEASURES
Version: PC90 (1990 IBM PC/MS-DOS) Date: 01/21/98 at 15:11:57

BETWEEN = EQUAL

EFFECT	VARIATE	STATISTIC	F	DF	P
A	SS = 0,031636	MS = 0,031636	0,93	1; 23	0,3444
B	SS = 0,527567	MS = 0,175856	5,18	3; 23	0,0070
C	SS = 0,033182	MS = 0,005530	0,16	6; 23	0,9841
ERROR	SS = 0,780930	MS = 0,033953			

BETWEEN = SIZES

EFFECT	VARIATE	STATISTIC	F	DF	P
A	SS = 0,101647	MS = 0,101647	2,99	1; 23	0,0970
B	SS = 0,551434	MS = 0,183811	5,41	3; 23	0,0058
C	SS = 0,033182	MS = 0,005530	0,16	6; 23	0,9841
ERROR	SS = 0,780930	MS = 0,033953			

4.2 Modelos com Estrutura do Tipo Fatorial-Hierárquica

Exemplo 2: “Esta pesquisa foi realizada pelo Centro de Pesquisa de Pecuária do Sudeste - CPPSE/EMBRAPA, São Carlos, SP, com o objetivo de estimar variâncias, covariâncias e parâmetros genéticos afins de características do desenvolvimento ponderal de animais

Canchim. Assim, tem-se os seguintes propósitos: (i) - obter estimadores de componentes de variâncias e covariâncias livres dos efeitos fixos; (ii) - conhecer os efeitos que interferem nos estimadores de parâmetros genéticos e (iii) - formar grupos de animais contemporâneos, ou seja, grupos de animais cuja variabilidade entre eles é devida ao acaso”.

Para os propósitos ilustrativos deste trabalho, considerou-se parte dos dados provenientes de bezerros da raça Canchim e reproduzidos na Tabela 13.

Tabela 13 - Peso à desmama, em kg, dos bezerros da raça Canchim

Fazenda		Sexo							
		Macho (A ₁)		Fêmea (A ₂)					
	Touro								
B ₁	C ₁	120	152	157	150	160	130	-	-
	C ₂	167	172	185	153	173	191	160	199
	C ₃	209	-	169	187	224	-	-	-
B ₂	C ₁	225	170	195	178	200	180	-	-
	C ₂	220	210	190	197	223	218	-	-

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

4.2.1 Obtenção das Somas de Quadrados

Considerando-se as diversas caracterizações assumidas pelos modelos (24) e (25) de 3.3.1 e 3.3.2, podem-se obter, por exemplo, os ajustes dos tipos sequenciais, (26) a (30) de 3.3.3, e que o SAS-GLM denota por tipo I. Assim, com base no conjunto de dados da Tabela 13, obtém-se as seguintes somas de quadrados ajustadas:

$$R(\mu_i | \mu) = R(\alpha | \mu) = 275,04173$$

$$R(\mu_i + \beta_j | \mu_i) = R(\beta | \mu, \alpha) = 3878,87792$$

$$R(\mu_{ij} | \mu_i + \beta_j) = R[(\alpha\beta) | \mu, \alpha, \beta] = 880,42650$$

$$R(\mu_{ij} + \gamma_{k(j)} | \mu_{ij}) = R[\gamma | \mu, \alpha, \beta, (\alpha\beta)] = 7116,22138$$

$$R(\mu_{ijk} | \mu_{ij} + \gamma_{k(j)}) = R[(\alpha\gamma) | \mu, \alpha, \beta, (\alpha\beta), \gamma] = 418,97093$$

No entanto, como já comentado em 3.4.2.1, por exemplo, se for considerada a ordem [B, A, A*B, C(B) e A*C(B)], a soma de quadrados do tipo I referente ao fator A, $R(\alpha | \mu, \beta)$, obtida tanto pelas parametrizações sucessivas quanto pelo SAS-GLM, não testa a hipótese do tipo I, pois ela depende da ordem de entrada dos fatores no modelo. Nesse caso, os usuários devem ficar atentos porque a soma de quadrados, $R(\alpha | \mu, \beta)$, não testa nem a hipótese do tipo I e nem a hipótese do tipo II.

Utilizando os procedimentos usuais, (32) a (36) de 3.3.3, obtém-se as seguintes somas de quadrados:

$$SQA = R(\alpha | \mu) = 5,33968$$

$$SQB = R(\beta | \mu) = 6746,68889$$

$$\begin{aligned} SQ(A*B) &= R[\mu, \alpha, \beta, (\alpha\beta)] - R(\mu, \alpha) - R(\mu, \beta) + R(\mu) \\ &= 433,12079 \end{aligned}$$

$$SQC(B) = R(\gamma | \mu, \beta) = 7978,52778$$

$$\begin{aligned} SQ[A*C(B)] &= R[\mu, \alpha, \beta, (\alpha\beta), \gamma, (\alpha\gamma)] - R[\mu, \alpha, \beta, (\alpha\beta)] - R(\mu, \beta, \gamma) + R(\mu, \beta) \\ &= 259,78953 \end{aligned}$$

Comparando-se esses resultados com os resultados da Tabela 14, verifica-se que à exceção da soma de quadrados do fator A, SQA, as demais somas de quadrados dos efeitos principais, obtidas através dos procedimentos usuais diferem daquelas fornecidas pelo SAS-GLM e, portanto, testam hipóteses totalmente distintas. Além disso, quando os dados são desbalanceados em presença ou não de caselas vazias, as hipóteses relacionadas às somas de quadrados $SQ(A*B)$ e $SQ[A*C(B)]$ não são associadas às hipóteses previamente definidas através das médias de caselas. Logo, elas testam hipóteses desconhecidas.

A fim de realizar a análise de variância através do SAS-GLM, utilizou-se o programa 6 e os resultados constam-se na Tabela 14.

Programa 6: Programa SAS-GLM para o modelo com estrutura fatorial-hierárquica

```

OPTIONS NODATE PS = 300;
DATA NESTFAT;
INPUT A B C Y;
CARDS;
1 1 1 120
1 1 1 152
1 1 2 167
.
.
.
2 2 2 223
2 2 2 218
;
PROC GLM;
CLASS A B C;
MODEL Y = A B A*B C(B) A*C(B) / SS1 SS2 SS3 SS4 E1 E2 E3 E4;
RUN;
```

Tabela 14 - 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	9	15423,4667	1713,7185	4,93	0,0014
Error	20	6954,0000	347,7000		
C. Total	29	22377,4667			
Source	DF	Type I	Type II	Type III	Type IV
A	1	5,33968	5,92111	<u>60,37051</u>	35,79085
B	1	6747,96474	6747,96474	5155,95947	5155,95947
A*B	1	431,84493	272,31629	154,79620	154,79620
C(B)	3	7818,30471	7818,30471	7768,74516	7768,74516
A*C(B)	3	420,01260	420,01260	420,01260	420,01260

4.2.2 Hipóteses Testadas Pelo SAS-GLM nos Modelos com Estrutura do Tipo Fatorial-Hierárquica

4.2.2.1 Hipóteses sobre o fator A

As hipóteses testadas pelas somas de quadrados, $R(\mu_i | \mu)$, $R(\alpha | \mu)$ e SQA, correspondem à hipótese do tipo I fornecida pelo SAS-GLM. Assim, elas testam, com 1 g.l., a hipótese sobre as médias ponderadas não ajustadas, dada por:

$$H_o^{(7)} : \frac{2\mu_{111} + 2\mu_{112} + \mu_{113} + 2\mu_{121} + 2\mu_{122}}{9} = \frac{4\mu_{211} + 6\mu_{212} + 3\mu_{213} + 4\mu_{221} + 4\mu_{222}}{21}$$

Descrevendo a hipótese $H_o^{(7)}$ na forma $H_o^{(7)} : B'\mu = \phi$ onde,

$$B' = [2/9 \ 2/9 \ 1/9 \ 2/9 \ 2/9 \ -4/21 \ -6/21 \ -3/21 \ -4/21 \ -4/21]$$

e
$$\mu = [\mu_{111} \ \mu_{112} \ \mu_{113} \ \mu_{121} \ \mu_{122} \ \mu_{211} \ \mu_{212} \ \mu_{213} \ \mu_{221} \ \mu_{222}] .$$

Então, a soma de quadrados associada à hipótese $H_o^{(7)}$ pode ser obtida através da estatística de Wald onde, das equações normais descrita em 3.2.1, tem-se:

$$\hat{\mu} = [136,0 \ 169,5 \ 209,0 \ 197,5 \ 215,0 \ 149,25 \ 176,83333 \ 193,33333 \ 188,25 \ 207,0] ;$$

$$(W'W)^{-1} = \text{diag} (1/ n_{ijk}) = \text{diag} (1/2, 1/2, 1, 1/2, 1/2, 1/4, 1/6, 1/3, 1/4, 1/4) ;$$

$$B'\hat{\mu} = 0,9206355 ; [B'(W'W)^{-1}B] = 0,1587302 \text{ e } [B'(W'W)^{-1}B]^{-1} = 6,3 .$$

Portanto,

$$SQH_o^{(7)} = 5,3396887 = R(\alpha | \mu) .$$

Como o SAS-GLM utiliza o modelo superparametrizado, nesse caso, substituindo-se μ_{ijk} por $\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(j)} + (\alpha\gamma)_{ik(j)}$; $H_o^{(7)}$ resulta em,

$$\begin{aligned} H_o^{(7)} : & \left\{ \alpha_1 + \frac{1}{63} \left(4\beta_2 + 35(\alpha\beta)_{11} + 28(\alpha\beta)_{12} + 2[\gamma_{1(1)} + \gamma_{1(2)} + \gamma_{2(2)}] + 14(\alpha\gamma)_{11(1)} + \right. \right. \\ & \left. \left. + 14(\alpha\gamma)_{12(1)} + 7(\alpha\gamma)_{13(1)} + 14(\alpha\gamma)_{11(2)} + 14(\alpha\gamma)_{12(2)} \right) \right\} = \\ & = \left\{ \alpha_2 + \frac{1}{63} \left(4\beta_1 + 39(\alpha\beta)_{21} + 24(\alpha\beta)_{22} + 2[2\gamma_{2(1)} + \gamma_{3(1)}] + 12(\alpha\gamma)_{21(1)} + \right. \right. \\ & \left. \left. + 18(\alpha\gamma)_{22(1)} + 9(\alpha\gamma)_{23(1)} + 12(\alpha\gamma)_{21(2)} + 12(\alpha\gamma)_{22(2)} \right) \right\} , \end{aligned}$$

como se observa na Tabela 15 fornecida pelo SAS-GLM, fazendo-se $L2=1$.

Tabela 15 - Funções estimáveis para o fator A fornecidas pelo SAS-GLM.

General Linear models Procedure					
Estimable Functions for: A					
Effects		Type I Coefficients	Type II Coefficients	Type III Coefficients	Type IV Coefficients
Intercept		0	0	0	0
A	1	L2	L2	L2	L2
	2	-L2	-L2	-L2	-L2
B	1	-0,0635*L2	0	0	0
	2	0,0635*L2	0	0	0
A*B	11	0,5556*L2	0,5733*L2	0,5294*L2	0,6*L2
	12	0,4444*L2	0,4267*L2	0,4706*L2	0,4*L2
	21	-0,619*L2	-0,5733*L2	-0,5294*L2	-0,6*L2
	22	-0,381*L2	-0,4267*L2	-0,4706*L2	-0,4*L2
C(B)	11	0,0317*L2	0	0	0
	21	-0,0635*L2	0	0	0
	31	-0,0317*L2	0	0	0
	12	0,0317*L2	0	0	0
	22	0,0317*L2	0	0	0
A*C(B)	111	0,2222*L2	0,2133*L2	0,1765*L2	0,2*L2
	121	0,2222*L2	0,24*L2	0,1765*L2	0,2*L2
	131	0,1111*L2	0,12*L2	0,1765*L2	0,2*L2
	211	-0,1905*L2	-0,2133*L2	-0,1765*L2	-0,2*L2
	221	-0,2857*L2	-0,24*L2	-0,1765*L2	-0,2*L2
	231	-0,1429*L2	-0,12*L2	-0,1765*L2	-0,2*L2
	112	0,2222*L2	0,2133*L2	0,2353*L2	0,2*L2
	122	0,2222*L2	0,2133*L2	0,2353*L2	0,2*L2
	212	-0,1905*L2	-0,2133*L2	-0,2353*L2	-0,2*L2
	222	-0,1905*L2	-0,2133*L2	-0,2353*L2	-0,2*L2

A hipótese testada pela soma de quadrados do tipo II, $R[\alpha | \mu, \beta, \gamma(\beta)]$, é diferente da hipótese testada pela soma de quadrados do tipo I, pois o conjunto de dados da Tabela 13 é desbalanceado. Nesse caso, testa a hipótese do tipo II, com 1 g.l., sobre as médias ponderadas ajustadas para os fatores B e C(B), dada por:

$$H_o^{(9)} : \begin{cases} \frac{2(4)}{6}(\mu_{111} - \mu_{211}) + \frac{2(6)}{8}(\mu_{112} - \mu_{212}) + \frac{1(3)}{4}(\mu_{113} - \mu_{213}) \\ + \frac{2(4)}{6}(\mu_{121} - \mu_{221}) + \frac{2(4)}{6}(\mu_{122} - \mu_{222}) = 0 \end{cases}$$

então,

$$B' = [8/6 \quad 12/8 \quad 3/4 \quad 8/6 \quad 8/6 \quad -8/6 \quad -12/8 \quad -3/4 \quad -8/6 \quad -8/6] ;$$

$$B' \hat{\mu} = 0,973333 ; [B'(W'W)^{-1}B] = 6,25 \text{ e } [B'(W'W)^{-1}B]^{-1} = 0,16 .$$

Nesse caso,

$$SQH_o^{(7)} = 5,9211111 = R [\alpha \mid \mu, \beta, \gamma(\beta)] .$$

Para obter a expressão equivalente àquela fornecida pelo SAS-GLM em sua saída, calculam-se as médias ponderadas de $H_o^{(9)}$, obtendo-se:

$$H_o^{(9)} : \begin{cases} \frac{16\mu_{111} + 18\mu_{112} + 9\mu_{113} + 16\mu_{121} + 16\mu_{122}}{75} = \\ = \frac{16\mu_{211} + 18\mu_{212} + 9\mu_{213} + 16\mu_{221} + 16\mu_{222}}{75} \end{cases}$$

onde, em termos de modelo superparametrizado, resulta em,

$$H_o^{(9)} : \left\{ \alpha_1 + \frac{1}{75} \left(43(\alpha\beta)_{11} + 32(\alpha\beta)_{12} + 16(\alpha\gamma)_{11(1)} + 18(\alpha\gamma)_{12(1)} + 9(\alpha\gamma)_{13(1)} + \right. \right. \\ \left. \left. + 16(\alpha\gamma)_{11(2)} + 16(\alpha\gamma)_{12(2)} \right) \right\} = \\ = \left\{ \alpha_2 + \frac{1}{75} \left(43(\alpha\beta)_{21} + 32(\alpha\beta)_{22} + 16(\alpha\gamma)_{21(1)} + 18(\alpha\gamma)_{22(1)} + 9(\alpha\gamma)_{23(1)} + \right. \right. \\ \left. \left. + 16(\alpha\gamma)_{21(2)} + 16(\alpha\gamma)_{22(2)} \right) \right\} ,$$

como se observa na Tabela 15 fazendo-se $L2 = 1$.

Para o conjunto de dados da Tabela 13, verifica-se que a soma de quadrados do tipo III referente ao fator A fornecida pelo SAS-GLM, não testa a hipótese sobre as médias não ponderadas, $H_o^{(10)}$, como pode ser observado na Tabela 15 fazendo-se $L2=1$. Nesse caso, a hipótese associada à soma de quadrados do tipo III, fornecida pelo SAS-GLM é da forma:

$$H_o^{(11)} : \frac{\frac{(\mu_{111} + \mu_{112} + \mu_{113})}{(3+1)} + \frac{(\mu_{121} + \mu_{122})}{(2+1)}}{\frac{3}{4} + \frac{2}{3}} = \frac{\frac{(\mu_{211} + \mu_{212} + \mu_{213})}{(3+1)} + \frac{(\mu_{221} + \mu_{222})}{(2+1)}}{\frac{3}{4} + \frac{2}{3}}$$

ou equivalentemente,

$$H_o^{(11)} : \frac{3(\mu_{111} + \mu_{112} + \mu_{113}) + 4(\mu_{121} + \mu_{122})}{17} = \frac{3(\mu_{211} + \mu_{212} + \mu_{213}) + 4(\mu_{221} + \mu_{222})}{17}$$

Usando a estatística de Wald, pode-se obter a soma de quadrados associada à hipótese $H_o^{(11)}$ onde,

$$B' = [3/17 \ 3/17 \ 3/17 \ 4/17 \ 4/17 \ -3/17 \ -3/17 \ -3/17 \ -4/17 \ -4/17] ;$$

$$B' \hat{\mu} = 3,191176 ; [B'(W'W)^{-1}B] = 0,168685 \text{ e } [B'(W'W)^{-1}B]^{-1} = 5,928205 .$$

Nesse caso,

$$SQH_o^{(11)} = 60,370513 .$$

Descrevendo a hipótese $H_o^{(11)}$ em termos de modelo superparametrizado como faz o SAS-GLM resulta em:

$$H_o^{(11)} : \left\{ \alpha_1 + 0.5294(\alpha\beta)_{11} + 0.4706(\alpha\beta)_{12} + 0.1765[(\alpha\gamma)_{11(1)} + (\alpha\gamma)_{12(1)} + (\alpha\gamma)_{13(1)}] + \right. \\ \left. + 0.2353[(\alpha\gamma)_{11(2)} + (\alpha\gamma)_{12(2)}] \right\} = \left\{ \alpha_2 + 0.5294(\alpha\beta)_{21} + 0.4706(\alpha\beta)_{22} + \right. \\ \left. + 0.1765[(\alpha\gamma)_{21(1)} + (\alpha\gamma)_{22(1)} + (\alpha\gamma)_{23(1)}] + 0.2353[(\alpha\gamma)_{21(2)} + (\alpha\gamma)_{22(2)}] \right\}$$

Novamente, verifica-se que a hipótese testada pela soma de quadrados do tipo III para o fator A, é gerada a partir de funções estimáveis complexas do tipo III de difícil interpretação. Nesse caso, considerando-se a hipótese testada, parece que nenhum procedimento estatístico usual é capaz de desenvolver essa hipótese. Em tais situações, seria desejável que as empresas fabricantes de *softwares* mostrassem claramente, através dos manuais, o que seus *softwares* estão calculando em suas saídas, a fim de que os usuários pudessem compreender e interpretar corretamente os resultados das análises. Caso contrário, a utilização de tais *softwares* mal documentados ou até mesmo mal elaborados, sem justificar o critério estatístico previamente definido pela literatura estatística, poderá comprometer seriamente os resultados obtidos após vários anos de pesquisa e a altos custos.

A soma de quadrados do tipo IV fornecida pelo SAS-GLM é equivalente ao método dos quadrados de médias ponderadas e testa a hipótese do tipo III sobre médias não ponderadas, dada por:

$$H_o^{(10)} : \frac{\mu_{111} + \mu_{112} + \mu_{113} + \mu_{121} + \mu_{122}}{5} = \frac{\mu_{211} + \mu_{212} + \mu_{213} + \mu_{221} + \mu_{222}}{5}$$

e então,

$$B' = [1/5 \ 1/5 \ 1/5 \ 1/5 \ 1/5 \ -1/5 \ -1/5 \ -1/5 \ -1/5 \ -1/5] ;$$

$$B'\hat{\mu} = 2,4666667 ; [B'(W'W)^{-1}B] = 0,17 \text{ e } [B'(W'W)^{-1}B]^{-1} = 5,8823529 .$$

Nesse caso, a soma de quadrados calculada através da estatística de Wald resulta em,

$$SQH_o^{(10)} = 35,79085.$$

Em termos de modelo superparametrizado, a hipótese $H_o^{(10)}$ fornecida pelo SAS-GLM é descrita na forma,

$$\begin{aligned} H_o^{(10)} : & \left\{ \alpha_1 + \frac{1}{5} \left(3(\alpha\beta)_{11} + 2(\alpha\beta)_{12} + (\alpha\gamma)_{11(1)} + (\alpha\gamma)_{12(1)} + (\alpha\gamma)_{13(1)} + \right. \right. \\ & \left. \left. + (\alpha\gamma)_{11(2)} + (\alpha\gamma)_{12(2)} \right) \right\} = \\ & = \left\{ \alpha_2 + \frac{1}{5} \left(3(\alpha\beta)_{21} + 2(\alpha\beta)_{22} + (\alpha\gamma)_{21(1)} + (\alpha\gamma)_{22(1)} + (\alpha\gamma)_{23(1)} + \right. \right. \\ & \left. \left. + (\alpha\gamma)_{21(2)} + (\alpha\gamma)_{22(2)} \right) \right\} , \end{aligned}$$

como pode ser observado na Tabela 15 fazendo-se $L2=1$.

4.2.2.2 Hipóteses sobre o fator B

Como os dados da Tabela 13 são desbalanceados, as somas de quadrados SQB e $R(\beta | \mu)$ que testam a hipótese do tipo I sobre as médias ponderadas não ajustadas, $H_o^{(13)}$, não são equivalentes às somas de quadrados $R(\beta | \mu, \alpha)$ e $R(\mu_i + \beta_j | \mu_i)$.

Porém, as somas de quadrados $R(\beta | \mu, \alpha)$ e $R(\mu_i + \beta_j | \mu_i)$ são equivalentes às somas de quadrados dos tipos I e II fornecidas pelo SAS-GLM. Nesse caso, testam a hipótese do tipo II sobre as médias ponderadas ajustadas:

$$H_o^{(12)} : \left\{ \frac{5(4)}{9} \left\{ \frac{2\mu_{111} + 2\mu_{112} + \mu_{113}}{5} - \frac{2\mu_{121} + 2\mu_{122}}{4} \right\} + \right. \\ \left. + \frac{13(8)}{21} \left\{ \frac{4\mu_{211} + 6\mu_{212} + 3\mu_{213}}{13} - \frac{4\mu_{221} + 4\mu_{222}}{8} \right\} \right\} = 0$$

onde, $B' = [8/9 \ 8/9 \ 4/9 \ -10/9 \ -10/9 \ 32/21 \ 48/21 \ 24/21 \ -52/21 \ -52/21]$;

$B'\hat{\mu} = -220,0317$; $[B'(W'W)^{-1}B] = 7,1746032$ e $[B'(W'W)^{-1}B]^{-1} = 0,1393805$.

Nesse caso, a soma de quadrados obtida através da estatística de Wald resulta em,

$$SQH_o^{(12)} = 6747,9647 = R(\beta | \mu, \alpha).$$

Calculando-se as médias ponderadas da expressão acima, obtém-se a expressão fornecida pelo SAS-GLM, dada por:

$$H_o^{(12)} : \left\{ \frac{56\mu_{111} + 56\mu_{112} + 28\mu_{113} + 96\mu_{211} + 144\mu_{212} + 72\mu_{213}}{452} = \right. \\ \left. = \frac{70\mu_{121} + 70\mu_{122} + 156\mu_{221} + 156\mu_{222}}{452} \right\}$$

onde, em termos do modelo superparametrizado resulta em,

$$H_o^{(12)} : \left\{ \alpha_1 + \frac{1}{452} \left(\begin{aligned} &140(\alpha\beta)_{11} + 312(\alpha\beta)_{21} + 152\gamma_{1(1)} + 200\gamma_{2(1)} + 100\gamma_{3(1)} + \\ &+ 56(\alpha\gamma)_{11(1)} + 56(\alpha\gamma)_{12(1)} + 28(\alpha\gamma)_{13(1)} + 96(\alpha\gamma)_{21(1)} + \\ &+ 144(\alpha\gamma)_{22(1)} + 72(\alpha\gamma)_{23(1)} \end{aligned} \right) \right\} =$$
$$= \left\{ \alpha_2 + \frac{1}{452} \left(\begin{aligned} &140(\alpha\beta)_{12} + 312(\alpha\beta)_{22} + 226\gamma_{1(2)} + 226\gamma_{2(2)} + 70(\alpha\gamma)_{11(2)} + \\ &+ 70(\alpha\gamma)_{12(2)} + 156(\alpha\gamma)_{21(2)} + 156(\alpha\gamma)_{22(2)} \end{aligned} \right) \right\},$$

como se observa na Tabela 16 fazendo-se L4=1.

Tabela 16 - Funções estimáveis para o fator B fornecidas pelo SAS-GLM.

General Linear Models Procedure			
Estimable Functions For: B			
Effects		Type I e II	Type III e IV
		Coefficients	Coefficients
Intercept		0	0
A	1	0	0
	2	0	0
B	1	L4	L4
	2	-L4	-L4
A*B	11	0,3097*L4	0,5*L4
	12	-0,3097L4	-0,5*L4
	21	0,6903*L4	0,5*L4
	22	-0,6903*L4	-0,5*L4
C(B)	11	0,3363*L4	0,3333*L4
	21	0,4425*L4	0,3333*L4
	31	0,2212*L4	0,3333*L4
	12	-0,5*L4	-0,5*L4
A*C(B)	22	-0,5*L4	-0,5*L4
	111	0,1239*L4	0,1667*L4
	121	0,1239*L4	0,1667*L4
	131	0,0619*L4	0,1667*L4
	211	0,2124*L4	0,1667*L4
	221	0,3186*L4	0,1667*L4
	231	0,1593*L4	0,1667*L4
	112	-0,1549*L4	-0,25*L4
	122	-0,1549*L4	-0,25*L4
	212	-0,3451*L4	-0,25*L4
	222	-0,3451*L4	-0,25*L4

As somas de quadrados dos tipos III e IV testam a hipótese do tipo III sobre as médias não ponderadas, dada por:

$$H_o^{(14)} : \frac{(\mu_{111} + \mu_{112} + \mu_{113} + \mu_{211} + \mu_{212} + \mu_{213})}{6} = \frac{(\mu_{121} + \mu_{122} + \mu_{221} + \mu_{222})}{4}.$$

Sendo assim, a soma de quadrados associada à hipótese $H_o^{(14)}$, pode ser obtida através da estatística de Wald onde,

$$B' = [1/6 \ 1/6 \ 1/6 \ -1/4 \ -1/4 \ 1/6 \ 1/6 \ 1/6 \ -1/4 \ -1/4] ;$$

$$B'\hat{\mu} = -29,61806 ; [B'(W'W)^{-1}B] = 0,1701389 \text{ e } [B'(W'W)^{-1}B]^{-1} = 5,877551 .$$

Logo,

$$SQH_o^{(14)} = 5155,9595.$$

Descrevendo a hipótese $H_o^{(14)}$ em termos de modelo superparametrizado como faz o SAS-GLM, obtém-se,

$$H_o^{(14)} : \left\{ \beta_1 + \frac{1}{2}[(\alpha\beta)_{11} + (\alpha\beta)_{21}] + \frac{1}{3}[\gamma_{1(1)} + \gamma_{2(1)} + \gamma_{3(1)}] + \frac{1}{6}[(\alpha\gamma)_{11(1)} + \right. \\ \left. + (\alpha\gamma)_{12(1)} + (\alpha\gamma)_{13(1)} + (\alpha\gamma)_{21(1)} + (\alpha\gamma)_{22(1)} + (\alpha\gamma)_{23(1)}] \right\} = \\ = \left\{ \beta_2 + \frac{1}{2}[(\alpha\beta)_{12} + (\alpha\beta)_{22}] + \frac{1}{2}[\gamma_{1(2)} + \gamma_{2(2)}] + \frac{1}{4}[(\alpha\gamma)_{11(2)} + (\alpha\gamma)_{12(2)} + \right. \\ \left. + (\alpha\gamma)_{21(2)} + (\alpha\gamma)_{22(2)}] \right\},$$

como se observa na Tabela 16, fornecida pelo sistema SAS-GLM, fazendo-se $L4=1$.

4.2.2.3 Hipóteses sobre a interação A*B

A hipótese testada pela soma de quadrados $SQ(A*B)$, obtida através do procedimento usual é diferente daquelas descritas em 3.4.2.3, pois é calculada por subtração. Porém, as somas de quadrados $R(\mu_{ij} | \mu_i + \beta_j)$ e $R[(\alpha\beta) | \mu, \alpha, \beta]$ são equivalentes à soma de quadrados do tipo I fornecida pelo SAS-GLM. Logo, elas testam a hipótese sobre a interação entre as médias ponderadas não ajustadas, dada por:

$$H_o^{(15)} : \left\{ \begin{array}{l} \frac{2\mu_{111} + 2\mu_{112} + \mu_{113}}{5} - \frac{4\mu_{211} + 6\mu_{212} + 3\mu_{213}}{13} - \frac{2\mu_{121} + 2\mu_{122}}{4} + \\ + \frac{4\mu_{221} + 4\mu_{222}}{8} = 0 \end{array} \right.$$

onde, $B' = [2/5 \ 2/5 \ 1/5 \ -2/4 \ -2/4 \ -4/13 \ -6/13 \ -3/13 \ 4/8 \ 4/8]$;

$$B'\hat{\mu} = -16,77885 ; [B'(W'W)^{-1}B] = 0,6519231 \text{ e } [B'(W'W)^{-1}B]^{-1} = 1,5339233.$$

Nesse caso,

$$SQH_o^{(15)} = 431,84492 = R[(\alpha\beta) | \mu, \alpha, \beta].$$

Descrevendo em termos de modelo superparametrizado, equivale a,

$$H_o^{(15)} : \left\{ \begin{array}{l} (\alpha\beta)_{11} + (\alpha\beta)_{22} + \frac{6}{65}\gamma_{1(1)} + \frac{1}{5}[2(\alpha\gamma)_{11(1)} + 2(\alpha\gamma)_{12(1)} + (\alpha\gamma)_{13(1)}] + \\ + \frac{1}{8}[4(\alpha\gamma)_{21(2)} + 4(\alpha\gamma)_{22(2)}] \end{array} \right\} =$$

$$= \left\{ \begin{array}{l} (\alpha\beta)_{12} + (\alpha\beta)_{21} + \frac{2}{65}(2\gamma_{2(1)} + \gamma_{3(1)}) + \frac{1}{13}[4(\alpha\gamma)_{21(1)} + 6(\alpha\gamma)_{22(1)} + \\ + 3(\alpha\gamma)_{23(1)}] + \frac{1}{4}[2(\alpha\gamma)_{11(2)} + 2(\alpha\gamma)_{12(2)}] \end{array} \right\},$$

como pode ser observado na Tabela 17, fazendo-se $L6=1$.

Tabela 17 - Funções estimáveis para a interação A*B
fornecidas pelo SAS-GLM.

General Linear Models Procedure				
Estimable Functions for: A*B				
Effects		Type I Coefficients	Type II Coefficients	Type III e IV Coefficients
Intercept		0	0	0
A	1	0	0	0
	2	0	0	0
B	1	0	0	0
	2	0	0	0
A*B	11	L6	L6	L6
	12	-L6	-L6	-L6
	21	-L6	-L6	-L6
	22	L6	L6	L6
C(B)	11	0,0923*L6	0	0
	21	-0,0615*L6	0	0
	31	-0,0308*L6	0	0
	12	0	0	0
	22	0	0	0
A*C(B)	111	0,4*L6	0,3721*L6	0,3333*L6
	121	0,4*L6	0,4186*L6	0,3333*L6
	131	0,2*L6	0,2093*L6	0,3333*L6
	211	-0,3077*L6	-0,3721*L6	-0,3333*L6
	221	-0,4615*L6	-0,4186*L6	-0,3333*L6
	231	-0,2308*L6	-0,2093*L6	-0,3333*L6
	112	-0,5*L6	-0,5*L6	-0,5*L6
	122	-0,5*L6	-0,5*L6	-0,5*L6
	212	0,5*L6	0,5*L6	0,5*L6
	222	0,5*L6	0,5*L6	0,5*L6

A soma de quadrados do tipo II fornecida pelo SAS-GLM testa a hipótese sobre a interação, de difícil interpretação, baseada em funções estimáveis complexas do tipo II, como se observa na Tabela 17 fazendo-se $L6=1$.

Já as somas de quadrados dos tipos III e IV testam hipóteses equivalentes

do tipo III sobre a interação entre as médias não ponderadas, dada por:

$$H_o^{(16)}: \frac{\mu_{111} + \mu_{112} + \mu_{113}}{3} - \frac{\mu_{211} + \mu_{212} + \mu_{213}}{3} - \frac{\mu_{121} + \mu_{122}}{2} + \frac{\mu_{221} + \mu_{222}}{2} = 0$$

onde,

$$B' = [1/3 \ 1/3 \ 1/3 \ -1/2 \ -1/2 \ -1/3 \ -1/3 \ -1/3 \ 1/2 \ 1/2] ;$$

$$B'\hat{\mu} = -10,26389 ; [B'(W'W)^{-1}B] = 0,6805556 \text{ e } [B'(W'W)^{-1}B]^{-1} = 1,4693878 .$$

Nesse caso,

$$SQH_o^{(16)} = 154,7962.$$

Descrevendo a hipótese $H_o^{(16)}$ em termos de modelo superparametrizado resulta em,

$$\begin{aligned} H_o^{(16)}: & \left\{ (\alpha\beta)_{11} + (\alpha\beta)_{22} + \frac{1}{3} [(\alpha\gamma)_{11(1)} + (\alpha\gamma)_{12(1)} + (\alpha\gamma)_{13(1)}] + \frac{1}{2} [(\alpha\gamma)_{21(2)} + (\alpha\gamma)_{22(2)}] \right\} \\ & = \left\{ (\alpha\beta)_{12} + (\alpha\beta)_{21} + \frac{1}{3} [(\alpha\gamma)_{21(1)} + (\alpha\gamma)_{22(1)} + (\alpha\gamma)_{23(1)}] + \frac{1}{2} [(\alpha\gamma)_{11(2)} + (\alpha\gamma)_{12(2)}] \right\} . \end{aligned}$$

4.2.2.4 Hipóteses sobre o fator C(B)

As somas de quadrados $R(\mu_{ij} + \gamma_{k(j)} \mid \mu_{ij})$ e $R[\gamma \mid \mu, \alpha, \beta, (\alpha\beta)]$, são equivalentes às somas de quadrados dos tipos I e II, fornecidas pelo SAS-GLM e, portanto, testam a hipótese conforme descrita em 3.4.2.4. Uma forma para a hipótese $H_o^{(17)}$, agora associada a 3 g.l., é dada por:

$$H_o^{(17)} : \begin{cases} (0.3023\mu_{111} - 0.0349\mu_{112} - 0.2674\mu_{113}) + \\ + (0.6977\mu_{211} + 0.0349\mu_{212} - 0.7326\mu_{213}) = 0 \\ \frac{2(1)}{3}(\mu_{112} - \mu_{113}) + \frac{6(3)}{9}(\mu_{212} - \mu_{213}) = 0 \\ \frac{2(2)}{4}(\mu_{121} - \mu_{122}) + \frac{4(4)}{8}(\mu_{221} - \mu_{222}) = 0 \end{cases}$$

onde, a primeira linha da hipótese $H_o^{(17)}$, é gerada a partir de funções estimáveis complexas do tipo II de difícil interpretação, pois o fator C(B) tem mais de dois níveis.

A soma de quadrados associada à hipótese $H_o^{(17)}$, pode ser obtida através da estatística de Wald onde,

$$B'_1 = [0,3023 \quad -0,0349 \quad -0,2674 \quad 0 \quad 0 \quad 0,6977 \quad 0,0349 \quad -0,7326 \quad 0 \quad 0];$$

$$B'_2 = [0 \quad 2/3 \quad -2/3 \quad 0 \quad 0 \quad 0 \quad 2 \quad -2 \quad 0 \quad 0];$$

$$B'_3 = [0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 2 \quad -2];$$

$$B' = [B'_1 | B'_2 | B'_3]'; \quad B' \hat{\mu} = [-52,02214 \quad -59,33333 \quad -55,0]';$$

$$[B'(W'W)^{-1}B] = \begin{bmatrix} 0,4186047 & 2/3 & 0 \\ & 8/3 & 0 \\ (\text{Sim.}) & & 3 \end{bmatrix};$$

$$e \quad [B'(W'W)^{-1}B]^{-1} = \begin{bmatrix} 3,9692307 & -0,992308 & 0 \\ & 0,6230769 & 0 \\ (\text{Sim.}) & & 1/3 \end{bmatrix}.$$

Sendo assim,

$$SQH_o^{(17)} = 7817,9757 \cong R[\gamma | \mu, \alpha, \beta, (\alpha\beta)].$$

Para obter as expressões equivalentes àquelas fornecidas pelo SAS-GLM,

calculam-se as médias ponderadas da segunda e terceira linhas de $H_o^{(17)}$, obtendo-se:

$$H_o^{(17)} : \begin{cases} 0.3023\mu_{111} + 0.6977\mu_{211} + 0.0349\mu_{212} = 0.0349\mu_{112} + 0.2674\mu_{113} + 0.7326\mu_{213} \\ \frac{2\mu_{112} + 6\mu_{212}}{8} = \frac{2\mu_{113} + 6\mu_{213}}{8} \\ \frac{\mu_{121} + 2\mu_{221}}{3} = \frac{\mu_{122} + 2\mu_{222}}{3} \end{cases}$$

que no modelo superparametrizado fornecida pelo SAS-GLM resulta em,

$$H_o^{(17)} : \begin{cases} \gamma_{1(1)} + 0,3023(\alpha\gamma)_{11(1)} + 0,6977(\alpha\gamma)_{21(1)} + 0,0349(\alpha\gamma)_{22(1)} = \\ = \gamma_{3(1)} + 0,0349(\alpha\gamma)_{12(1)} + 0,2674(\alpha\gamma)_{13(1)} + 0,7326(\alpha\gamma)_{23(1)} \\ \gamma_{2(1)} + \frac{2(\alpha\gamma)_{12(1)} + 6(\alpha\gamma)_{22(1)}}{8} = \gamma_{3(1)} + \frac{2(\alpha\gamma)_{13(1)} + 6(\alpha\gamma)_{23(1)}}{8} \\ \gamma_{1(2)} + \frac{(\alpha\gamma)_{11(2)} + 2(\alpha\gamma)_{21(2)}}{3} = \gamma_{2(2)} + \frac{(\alpha\gamma)_{12(2)} + 2(\alpha\gamma)_{22(2)}}{3} \end{cases}$$

como se observa na Tabela 18, fazendo-se $L10=1$ e $L11=L13=0$; $L11=1$ e $L10=L13=0$; $L10=L11=0$ e $L13=1$, pois o fator C(B) tem 3 g.l..

As somas de quadrados dos tipos III e IV fornecidas pelo SAS-GLM testam hipóteses equivalentes do tipo III sobre as médias não ponderadas. Uma forma para a hipótese $H_o^{(18)}$ é:

$$H_o^{(18)} : \begin{cases} \frac{\mu_{111} + \mu_{211}}{2} + \frac{\mu_{112} + \mu_{212}}{2} = \mu_{113} + \mu_{213} \\ \frac{\mu_{112} + \mu_{212}}{2} = \frac{\mu_{113} + \mu_{213}}{2} \\ \frac{\mu_{121} + \mu_{221}}{2} = \frac{\mu_{122} + \mu_{222}}{2} \end{cases}$$

onde,

$$B'_1 = [1/2 \ 1/2 \ -1 \ 0 \ 0 \ 1/2 \ 1/2 \ -1 \ 0 \ 0];$$

$$B'_2 = [0 \ 1/2 \ -1/2 \ 0 \ 0 \ 0 \ 1/2 \ -1/2 \ 0 \ 0];$$

$$B'_3 = [0 \ 0 \ 0 \ 1/2 \ -1/2 \ 0 \ 0 \ 0 \ 1/2 \ -1/2];$$

$$B' = [B'_1 | B'_2 | B'_3]'; \quad B'\hat{\mu} = [-86,54167 \ -28,0 \ -18,125]';$$

$$[B'(W'W)^{-1}B] = \begin{bmatrix} 27/16 & 5/6 & 0 \\ & 1/2 & 0 \\ (\text{Sim.}) & & 3/8 \end{bmatrix};$$

$$\text{e } [B'(W'W)^{-1}B]^{-1} = \begin{bmatrix} 144/43 & -240/43 & 0 \\ & 486/43 & 0 \\ (\text{Sim.}) & & 8/3 \end{bmatrix}.$$

Logo,

$$SQH_o^{(18)} = 7768,7452.$$

Descrevendo em termos de modelo superparametrizado, como faz o SAS-

GLM, obtém-se:

$$H_o^{(18)} : \begin{cases} \gamma_{1(1)} + \gamma_{2(1)} + \frac{(\alpha\gamma)_{11(1)} + (\alpha\gamma)_{21(1)}}{2} + \frac{(\alpha\gamma)_{12(1)} + (\alpha\gamma)_{22(1)}}{2} = 2\gamma_{3(1)} + (\alpha\gamma)_{13(1)} + \\ + (\alpha\gamma)_{23(1)} \\ \gamma_{2(1)} + \frac{(\alpha\gamma)_{12(1)} + (\alpha\gamma)_{22(1)}}{2} = \gamma_{3(1)} + \frac{(\alpha\gamma)_{13(1)} + (\alpha\gamma)_{23(1)}}{2} \\ \gamma_{1(2)} + \frac{(\alpha\gamma)_{11(2)} + (\alpha\gamma)_{21(2)}}{2} = \gamma_{2(2)} + \frac{(\alpha\gamma)_{12(2)} + (\alpha\gamma)_{22(2)}}{2} \end{cases}$$

como se observa na Tabela 18, fazendo-se $L_{10}=L_{11}=1$ e $L_{13}=0$; $L_{11}=1$ e $L_{10}=L_{13}=0$; $L_{10}=L_{11}=0$ e $L_{13}=1$, onde cada L representa o número de graus de liberdade do fator C(B).

Tabela 18 - Funções estimáveis para o fator C(B) e interação A*C(B) fornecidas pelo SAS-GLM.

General Linear Models Procedure				
Estimable Functions for:				
		C(B)		A*C(B)
		Type I e II	Type III e IV	Type I, II, III, IV
Effects		Coefficients	Coefficients	Coefficients
Intercept		0	0	0
A	1	0	0	0
	2	0	0	0
B	1	0	0	0
	2	0	0	0
A*B	11	0	0	0
	12	0	0	0
	21	0	0	0
	22	0	0	0
C(B)	11	L10	L10	0
	21	L11	L11	0
	31	-L10-L11	-L10-L11	0
	12	L13	L13	0
	22	-L13	-L13	0
A*C(B)	111	0,3023*L10	0,5*L10	L15
	121	-0,0349*L10+0,25*L11	0,5*L11	L16
	131	-0,2674*L10-0,25*L11	-0,5*L10-0,5*L11	-L15-L16
	211	0,6977*L10	0,5*L10	-L15
	221	0,0349*L10+0,75*L11	0,5*L11	-L16
	231	-0,7326*L10-0,75*L11	-0,5*L10-0,5*L11	L15+L16
	112	0,3333*L13	0,5*L13	L21
	122	-0,3333*L13	-0,5*L13	-L21
	212	0,6667*L13	0,5*L13	-L21
	222	-0,6667*L13	-0,5*L13	L21

4.2.2.5 Hipótese sobre a interação A*C(B)

Para a interação A*C(B), a hipótese testada pela soma de quadrados $SQ[A*C(B)]$, obtida através do procedimento usual não é definida, pois ela também é calculada por subtração. Já as somas de quadrados dos tipos I, II, III e IV fornecidas pelo SAS-GLM, bem como aquelas obtidas através da redução nas somas de quadrados, $R(\mu_{ijk} | \mu_{ij} + \gamma_{k(j)})$ e $R[(\alpha\gamma) | \mu, \alpha, \beta, (\alpha\beta), \gamma]$, testam hipóteses equivalentes, $H_o^{(20)}$, com 3 g.l., conforme descrita em 3.4.2.5. Nesse caso, é dada por:

$$H_o^{(20)} : \begin{cases} \mu_{111} - \mu_{113} - \mu_{211} + \mu_{213} = 0 \\ \mu_{112} - \mu_{113} - \mu_{212} + \mu_{213} = 0 \\ \mu_{121} - \mu_{122} - \mu_{221} + \mu_{222} = 0 \end{cases}$$

então,

$$B'_1 = [1 \ 0 \ -1 \ 0 \ 0 \ -1 \ 0 \ 1 \ 0 \ 0];$$

$$B'_2 = [0 \ 1 \ -1 \ 0 \ 0 \ 0 \ -1 \ 1 \ 0 \ 0];$$

$$B'_3 = [0 \ 0 \ 0 \ 1 \ -1 \ 0 \ 0 \ 0 \ -1 \ 1];$$

$$B' = [B'_1 | B'_2 | B'_3]'; \quad B'\hat{\mu} = [-28,91667 \quad -23,0 \quad 1,25]';$$

$$[B'(W'W)^{-1}B] = \begin{bmatrix} 25/12 & 4/3 & 0 \\ & 2 & 0 \\ (\text{Sim.}) & & 3/2 \end{bmatrix};$$

$$\text{e } [B'(W'W)^{-1}B]^{-1} = \begin{bmatrix} 36/43 & -24/43 & 0 \\ & 75/86 & 0 \\ (\text{Sim.}) & & 2/3 \end{bmatrix}.$$

Portanto,

$$SQH_o^{(20)} = 420,0126 = R [(\alpha\gamma) \mid \mu, \alpha, \beta, \gamma, (\alpha\beta)].$$

Descrevendo em termos de modelo superparametrizado tem-se,

$$H_o^{(20)} : \begin{cases} (\alpha\gamma)_{11(1)} - (\alpha\gamma)_{13(1)} - (\alpha\gamma)_{21(1)} + (\alpha\gamma)_{23(1)} = 0 \\ (\alpha\gamma)_{12(1)} - (\alpha\gamma)_{13(1)} - (\alpha\gamma)_{22(1)} + (\alpha\gamma)_{23(1)} = 0 \\ (\alpha\gamma)_{11(2)} - (\alpha\gamma)_{12(2)} - (\alpha\gamma)_{21(2)} + (\alpha\gamma)_{22(2)} = 0 \end{cases}$$

como observa-se na Tabela 18, fazendo-se $L16=L21=0$ e $L15=1$; $L15=L21=0$ e $L16=1$; $L15=L16=0$ e $L21=1$, pois a interação $A*C(B)$ possui 3 g.l..

Como já comentado, tanto no modelo com estrutura do tipo estritamente hierárquica como no modelo com estrutura do tipo fatorial-hierárquica, as somas de quadrados do tipo III, referentes ao fator A fornecidas pelo SAS-GLM não testam hipóteses do tipo III sobre as médias não ponderadas, $H_o^{(2)}$ e $H_o^{(10)}$, mas hipóteses completamente diferentes, $H_o^{(3)}$ e $H_o^{(11)}$, baseadas em funções estimáveis complexas do tipo III.

A fim de compreender e tentar explicar os resultados dessas somas de quadrados do tipo III, que não testam as hipóteses do tipo III, Searle (1994) discutiu três razões para o exemplo considerado por Dallal (1992), que são enfatizadas a seguir.

- Searle (1994) considerou que uma das razões fosse devida à ordem de entrada dos fatores principais nos modelos. Mas, como constatado nesse trabalho, apenas as somas de quadrados do tipo I são afetadas pela sequência de entrada dos fatores. As somas

de quadrados do tipo III não são afetadas, pois elas são ajustadas para todos os fatores e interações.

- Segundo o autor, uma outra possível razão para o problema, talvez estivesse relacionada às funções estimáveis que geram as somas de quadrados do tipo III. Dessa forma, esperava que se pudesse estabelecer uma explicação para a hipótese do tipo III e, portanto, compreender o comportamento das somas de quadrados do tipo III. Mas infelizmente, também não foi possível. Isto porque, embora as hipóteses $H_o^{(3)}$ e $H_o^{(11)}$ tenham sido desenvolvidas aqui, elas são derivadas a partir de funções estimáveis complexas. Isto levou a considerar uma outra razão para a soma de quadrados do tipo III referente ao fator A não testasse a hipótese $H_o^{(2)}$ ou $H_o^{(10)}$.

- O autor acredita que, essa diferença seja devida ao desbalanceamento dos dados com relação aos níveis dos fatores. Por exemplo, observando-se os dados da Tabela 13, pode ser visto que dentro de cada nível do fator B existe balanceamento com relação aos níveis de C(B). Assim, em B1 existem três níveis do fator C(B) para cada nível do fator A, e em B2 existem dois níveis do fator C(B) para cada nível de A. Aparentemente, por causa desse aspecto de balanceamento, a soma de quadrados do tipo III para o fator B, testa realmente a hipótese do tipo III, aqui denotada por $H_o^{(14)}$. Ao contrário, essa forma de balanceamento não se aplica para a soma de quadrados do tipo III do fator A. Dentro de cada nível do fator A não existe balanceamento com relação aos níveis do fator C(B) sobre os níveis do fator B. Assim, verifica-se que dentro do fator A existem três níveis do fator C(B) em B1, mas apenas dois níveis do fator C(B) em B2. Aparentemente, isto fez com que a soma de quadrados tipo III não testasse a hipótese tipo III, denotada por

$H_o^{(10)}$

Para comprovar o argumento de Searle (1994) sobre desbalanceamento dos dados com relação aos níveis do fator C(B), foram retiradas as observações y_{2221} , y_{2222} , y_{2223} e y_{2224} , de modo que o fator B também ficasse desbalanceado.

Para essa configuração, calculou-se novamente as somas de quadrados dos tipos I, II, III e IV. Constatou-se que (Tabela 19), agora as somas de quadrados do tipo III para os fatores A e B não testaram as hipóteses do tipo III, $H_o^{(10)}$ e $H_o^{(14)}$, mas hipóteses completamente diferentes, como podem ser observadas na Tabela 21, fazendo-se $L2=1$ para os efeitos do fator A e $L4=1$ para os efeitos do fator B.

Tabela 19 - 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	8	12569,5385	1571,1923	4,32	0,0054
Error	17	6188,0000	364,0000		
C. Total	25	18757,5385			
Source	DF	Type I SS	Type II SS	Type III SS	Type IV SS
A	1	275,04173	4,27260	<u>32,08816</u>	5,36508*
B	1	3878,87792	3878,87792	<u>3495,61336</u>	2644,53704
A*B	1	880,42650	189,67314	112,32749	112,32749*
C(B)	3	7116,22138	7116,22138	7198,95349	7198,95349*
A*C(B)	2	481,97093	481,97093	418,97093	418,97093

* Note: Other Type IV Testable Hypotheses Exist Which May Yield Different SS.

Tabela 20 - Hipóteses associadas às somas de quadrados fornecidas pelo SAS-GLM

Source	Type I SS	Type II SS	Type III SS	Type IV SS
A	$H_o^{(7)}$	$H_o^{(9)*}$	$H_o^{(11)*}$	$H_o^{(10)*}$
B	$H_o^{(12)}$	$H_o^{(12)}$	Complexas	Complexas
A*B	$H_o^{(15)}$	Complexas	$H_o^{(16)*}$	$H_o^{(16)*}$
C(B)	$H_o^{(17)**}$	$H_o^{(17)**}$	$H_o^{(18)**}$	$H_o^{(18)**}$
A*C(B)	$H_o^{(20)***}$	$H_o^{(20)***}$	$H_o^{(20)***}$	$H_o^{(20)***}$

* - Hipóteses sem considerar a média μ_{122}
** - Hipóteses sem considerar a média μ_{221}
*** - Hipótese sem considerar as médias μ_{121} , μ_{122} e μ_{221}

Observa-se na Tabela 19 que a soma de quadrados do tipo III referente ao fator A, gerada a partir de funções estimáveis complexas do tipo III, testa a hipótese $H_o^{(11)}$ apenas para as médias de caselas das linhas que estão completas, pois o fator A tem somente dois níveis.

$$H_o^{(11)} : \frac{\frac{(\mu_{111} + \mu_{112} + \mu_{113})}{(3 + 1)} + \frac{\mu_{121}}{(1 + 1)}}{\frac{3}{4} + \frac{1}{2}} = \frac{\frac{(\mu_{211} + \mu_{212} + \mu_{213})}{(3 + 1)} + \frac{\mu_{221}}{(1 + 1)}}{\frac{3}{4} + \frac{1}{2}}$$

ou equivalentemente,

$$H_o^{(11)} : \frac{\mu_{111} + \mu_{112} + \mu_{113} + 2\mu_{121}}{5} = \frac{\mu_{211} + \mu_{212} + \mu_{213} + 2\mu_{221}}{5}$$

onde, $B' = [1/5 \ 1/5 \ 1/5 \ 2/5 \ -1/5 \ -1/5 \ -1/5 \ -2/5]$
e $\hat{\mu} = [136,0 \ 169,5 \ 209,0 \ 197,5 \ 149,25 \ 176,83333 \ 193,33333 \ 188,25]$.

Logo, a soma de quadrados associada à hipótese $H_0^{(11)}$ pode ser obtida através da estatística de Wald,

$$(W'W)^{-1} = \text{diag} (1/ n_{ijk}) = \text{diag} (1/2, 1/2, 1, 1/2, 1/4, 1/6, 1/3, 1/4) ;$$

$$B'\hat{\mu} = 2,7166667 ; [B'(W'W)^{-1}B] = 0,23 \text{ e } [B'(W'W)^{-1}B]^{-1} = 4,3478261 .$$

Portanto, a soma de quadrados resulta em,

$$SQH_0^{(11)} = 32,088164 .$$

A hipótese referente ao fator B testada pelo SAS-GLM, também baseia-se em funções estimáveis complexas do tipo III de difícil interpretação.

$$H_0^{(22)} : \frac{349(\mu_{111} + \mu_{112} + \mu_{113}) + 276(\mu_{211} + \mu_{212} + \mu_{213})}{1875} = \frac{211\mu_{121} + 487\mu_{122} + 552\mu_{221}}{1250}$$

onde,

$$B' = \begin{bmatrix} 349/1875 & 349/1875 & 349/1875 & -211/1250 & -487/1250 & 276/1875 \\ 276/1875 & 276/1875 & -552/1250 & & & \end{bmatrix}$$

$$\hat{\mu} = [136,0 \ 169,5 \ 209,0 \ 197,5 \ 215,0 \ 149,25 \ 176,83333 \ 193,33333 \ 188,25] ;$$

$$(W'W)^{-1} = \text{diag} (1/ n_{ijk}) = \text{diag} (1/2, 1/2, 1, 1/2, 1/2, 1/4, 1/6, 1/3, 1/4) ;$$

$$B'\hat{\mu} = -28,00947 ; [B'(W'W)^{-1}B] = 0,2244356 \text{ e } [B'(W'W)^{-1}B]^{-1} = 4,455622 .$$

Nesse caso, a soma de quadrados obtida através da estatística de Wald resulta em,

$$SQH_0^{(**)} = 3495,5702 .$$

Tabela 21 - Funções estimáveis para os fatores A e B
fornecidas pelo SAS-GLM:

General Linear Models Procedure			
Type III Estimable Functions for:			
		A	B
Effects		Coefficients	Coefficients
Intercept		0	0
A	1	L2	0
	2	-L2	0
B	1	0	L4
	2	0	-L4
A*B	11	0,6*L2	0,5584*L4
	12	0,4*L2	-0,5584*L4
	21	-0,6*L2	0,4416*L4
	22	-0,4*L2	-0,4416*L4
C(B)	11	0	0,3333*L4
	21	0	0,3333*L4
	31	0	0,3333*L4
	12	0	-0,6104*L4
	22	0	-0,3896*L4
A*C(B)	111	0,2*L2	0,1861*L4
	121	0,2*L2	0,1861*L4
	131	0,2*L2	0,1861*L4
	211	-0,2*L2	0,1472*L4
	221	-0,2*L2	0,1472*L4
	231	-0,2*L2	0,1472*L4
	112	0,4*L2	-0,1688*L4
	122	0	-0,3896*L4
	212	-0,4*L2	-0,4416*L4

Assim, os resultados indicam que não existe uma explicação estatística lógica para o SAS-GLM testar essas hipóteses, conforme constatado por Searle (1994) e, provavelmente, são hipóteses sem interesse científico para os pesquisadores. Desse modo, sem uma explicação estatística detalhada da razão dessas funções estimáveis

complexas do tipo III serem geradas dessa forma, e levando em consideração as hipóteses que elas testam, tudo indica que nenhum procedimento estatístico usual é capaz de desenvolver ou testar essas hipóteses.

Nesse contexto, talvez seja mais prudente não utilizar as somas de quadrados do tipo III em modelos de efeitos fixos que envolvam fatores cruzados e hierárquicos com dados desbalanceados.

4.2.3 Hipóteses Testadas por Outros *Softwares* nos Modelos com Estrutura do Tipo Fatorial-Hierárquica

De modo análogo ao do exemplo 1, a fim de comparar as hipóteses testadas pelo SAS-GLM com outros *Softwares*, foram utilizados o MINITAB, o NTIA, o STATGRAPHICS, o SAEG, o SANEST, o GLIM, o STATISTICA e o BMDP.

4.2.3.1 MINITAB

Programa 7: Programa MINITAB para modelos com Estrutura Fatorial-Hierárquica

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

Quando os dados são desbalanceados com todas as caselas ocupadas, o procedimento GLM do MINITAB fornece somas de quadrados dos tipos sequenciais e ajustadas, equivalentes às somas de quadrados dos tipos I e III fornecidas pelo SAS-

GLM. Porém, no modelo com estrutura do tipo fatorial-hierárquica, o procedimento GLM do MINITAB fornece apenas somas de quadrados seqüenciais e não realiza nenhum teste estatístico, pois os níveis do fator C(B) não são iguais.

Tabela 22 - Análise de variância fornecida pelo MINITAB

General Linear Model			
Analysis of Variance for Y			
Source	Model DF	Reduced DF	Seq. SS
A	1	1	5,3
B	1	1	6748,0
A*B	1	1	431,8
C(B)	4	3+	7818,3
A*C(B)	4	3+	420,0
Error	18	20	6954,0
Total	29	29	22377,5

+ Rank deficiency due to empty cells, unbalanced nesting or collinearity. No storage of results or further analysis will be done.

4.2.3.2 NTIA

Programa 8: Programa NTIA para modelos com Estrutura Fatorial-Hierárquica

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

O NTIA fornece somas de quadrados dos tipos seqüenciais e parciais, equivalentes às somas de quadrados dos tipos I e III fornecidas pelo SAS-GLM, exceto para a soma de quadrados do tipo III referente ao fator A que, nesse caso, forneceu $SQA = 71,71456916$ e, portanto, não testa a hipótese do tipo III sobre as médias não ponderadas, $H_0^{(10)}$. A soma de quadrados fornecida pelo NTIA é calculada através da notação-R (.) , $R [\alpha^* | \mu^*, \beta^*, (\alpha\beta)^*, \gamma^*, (\alpha\gamma)^*]$, utilizando-se o modelo reparametrizado de posto completo e, nesse caso, não é apropriada, pois os níveis do fator C(B) não são os mesmos dentro de cada nível do fator B.

Tabela 23 - Análise de variância fornecida pelo NTIA

F. de variação	G.L	S.Q Seq.	Q.M	Valor F	PR > F
A	1	5,33968254	5,33968254	0,0154	0,903
B	1	6747,96474224	6747,96474224	19,4074	0,000
A*B	1	431,84493420	431,84493420	1,2420	0,278
C(B)	3	7818,30471079	2606,10157026	7,4953	0,001
A*C(B)	3	420,01259690	140,00419897	0,4027	0,753
Resíduo	20	6954,00000000	347,70000000		
Total	29	22377,46666667			
F. de variação	G.L	S.Q Parciais	Q.M	Valor F	PR > F
A	1	<u>71,71456916</u>	71,71456916	0,2063	0,655
B	1	5155,95946712	5155,95946712	14,8288	0,001
A*B	1	154,79620181	154,79620181	0,4452	0,512
C(B)	3	7768,74515504	2589,58171835	7,4477	0,002
A*C(B)	3	420,01259690	140,00419897	0,4027	0,753

4.2.3.3 STATGRAPHICS

O STATGRAPHICS não realiza as análises para modelos com estrutura do tipo fatorial-hierárquica.

4.2.3.4 SAEG

O SAEG também não realiza as análises para modelos com estrutura do tipo fatorial-hierárquica.

4.2.3.5 SANEST

O SANEST não realiza as análises para modelos com estrutura do tipo fatorial-hierárquica.

4.2.3.6 GLIM

Programa 9: Programa GLIM para modelos com Estrutura Fatorial-Hierárquica

```
$UNITS 30 $DATA A B C Y $READ $!
```

```
1 1 1 120
```

```
1 1 1 152
```

```
1 1 2 167
```

```
.
```

```
.
```

```
.
```

```
2 2 2 223
```

```
2 2 2 218
```

```
$FACTOR A 2 B 2 C 3 $YVAR Y $FIT: +A: +B:+A.B: +C/B: +A.C/B $!
```

```
$FINISH $!
```

O GLIM também fornece apenas somas de quadrados seqüenciais equivalentes às somas de quadrados do tipo I fornecidas pelo SAS-GLM, sem realizar nenhum teste estatístico. Como se observa, os resultados fornecidos pelo GLIM são extremamente difíceis de interpretar, pois a sua saída não evidencia os efeitos envolvidos.

Tabela 24 - Análise de variância fornecida pelo GLIM

[o]	GLIM 4, update 8 for IBM etc. 80386 PC / DOS on 21-Jan-1998
[o]	(copyright) 1992 Royal Statistical Society, London
[o]	
[o]	deviance = 22377
[o]	residual df = 29
[o]	
[o]	deviance = 22372 (change = -5,340)
[o]	residual df = 28 (change = -1)
[o]	
[o]	deviance = 15624 (change = -6748)
[o]	residual df = 27 (change = -1)
[o]	
[o]	deviance = 15192 (change = -431,8)
[o]	residual df = 26 (change = -1)
[o]	
[o]	deviance = 7374 (change = -7818)
[o]	residual df = 23 (change = -3)
[o]	
[o]	deviance = 6954 (change = -420)
[o]	residual df = 20 (change = -3)

4.2.3.7 STATISTICA

Tabela 25 - Comparações planejados para os efeitos principais e interação
A*C(B) fornecidos pelo STATISTICA

Planned Comparison - Fator A					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	35,791	1	35,7908	,102936	,751663
Error	6954,000	20	347,7000		
Planned Comparison - Fator B					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	5155,959	1	5155,959	14,82876	,000997
Error	6954,000	20	347,700		
Planned Comparison - Fator C(B)					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	7768,745	3	2589,582	7,447747	,001544
Error	6954,000	20	347,700		
Planned Comparison - Interação A*C(B)					
1-VAR1, 2-VAR2, 3-VAR3					
	SS	df	MS	F	p-level
Effect	420,013	3	140,0042	,402658	,752649
Error	6954,000	20	347,7000		

Quando os dados são desbalanceados, o STATISTICA não realiza as análises para modelos com estrutura do tipo fatorial-hierárquica. Porém, através do comando

“CONTRASTS FOR BETWEEN-GROUP FACTORS”, pode-se obter as somas de quadrados dos efeitos principais A, B, C(B) e da interação A*C(B). Nesse caso, se todas as caselas estão ocupadas, elas são equivalentes às somas de quadrados do tipo IV fornecidas pelo SAS-GLM, portanto, testam as hipóteses $H_0^{(10)}$, $H_0^{(14)}$, $H_0^{(18)}$ e $H_0^{(20)}$, respectivamente.

4.2.3.8 BMDP

Programa 10: Programa GLIM para modelos com Estrutura Fatorial-Hierárquica

```

/ INPUT   TITLE IS 'MODELO FATORIAL-HIERARQUICOS'.
          VARIABLES = 4.
          FORMAT = FREE.
/ VARIABLE NAMES = A, B, C, Y.
/ BETWEEN FACTORS = A, B, C.
          CODES(A) = 1, 2.
          CODES(B) = 1, 2.
          CODES(C) = 1 TO 5.
/ WEIGHTS BETWEEN = EQUAL.
/END
1 1 1 120
1 1 1 152
.
.
2 2 5 223
2 2 5 218
/END
ANALYSIS  PROCEDURE = STRUCTURE.
          BFORMULA = 'A*B/C'./
END /
/ WEIGHT  BETWEEN = SIZES.
/END
ANALYSIS  PROCEDURE = STRUCTURE.
          BFORMULA = 'A*B/C'./
END /

```

Tabela 26 - Análise de variância fornecida pelo BMDP

BMDP4V—UNIVARIATE AND MULTIVARIATE ANALYSIS OF VARIANCE AND COVARIANCE, INCLUDING REPEATED MEASURES

Version: PC90 (1990 IBM PC/MS-DOS) Date: 01/21/98 at 08:53:45

BETWEEN = EQUAL

EFFECT	VARIATE	STATISTIC	F	D.F	P
A	SS = 35,79085	MS = 35,79085	0,10	1; 20	0,7517
B	SS = 5155,95947	MS = 5155,95947	14,83	1; 20	0,0010
A*B	SS = 154,79620	MS = 154,79620	0,45	1; 20	0,5122
C	SS = 7768,74515	MS = 2589,58172	7,45	3; 20	0,0015
AC	SS = 420,01260	MS = 140,00420	0,40	3; 20	0,7526
ERROR	SS = 6954,00000	MS = 347,70000			

BETWEEN = SIZES

EFFECT	VARIATE	STATISTIC	F	D.F	P
A	SS = 5,33968	MS = 5,33968	0,02	1; 20	0,9026
B	SS = 6746,68889	MS = 6746,68889	19,40	1; 20	0,0003
A*B	SS = 431,84493	MS = 431,84493	1,24	1; 20	0,2783
C	SS = 7978,52778	MS = 2659,50926	7,65	3; 20	0,0013
AC	SS = 420,01256	MS = 140,00420	0,40	3; 20	0,7526
A B,AB,C	SS = 5,61845	MS = 5,61845	0,02	1; 20	0,9001
B A,AB,C	SS = 6699,66864	MS = 6699,66864	19,27	1; 20	0,0003
AB A,B,C	SS = 272,31629	MS = 272,31629	0,78	1; 20	0,3867
C A,B,AB	SS = 7818,30471	MS = 2606,10157	7,50	3; 20	0,0015

Como os níveis do fator C(B) são diferentes dentro de cada nível do fator

B, o comando BETWEEN = EQUAL do BMDP fornece somas de quadrados equivalentes às somas de quadrados do tipo IV fornecidas pelo SAS-GLM.

Já o comando BETWEEN = SIZES do BMDP fornece dois quadros para a análise da variância. No primeiro quadro, todas as somas de quadrados fornecidas pelo BMDP testam as hipóteses do tipo I sobre as médias ponderadas não ajustadas. Como o SAS-GLM, faz ajustes do tipo seqüencial para as somas de quadrados do tipo I, nesse caso, à exceção da soma de quadrados do fator A, as demais somas de quadrados do BMDP são diferentes das somas de quadrados do tipo I fornecidas pelo SAS-GLM. No segundo quadro, as somas de quadrados dos fatores A e B fornecidas pelo BMDP não são equivalentes às somas de quadrados do tipo II fornecidas pelo SAS-GLM e, portanto, não testam as hipóteses do tipo II. As somas de quadrados da interação A*B e do fator C(B) são equivalentes às somas de quadrados do tipo II do SAS-GLM.

Para o caso de dados desbalanceados com casela vazia, onde foram retiradas as observações y_{2221} , y_{2222} , y_{2223} e y_{2224} , o SANEST, o SAEG e o STATGRAPHICS não realizaram as análises. O MINITAB, o GLIM e inclusive o NTIA que não calcula somas de quadrados parciais quando existe caselas vazias, fornecem apenas somas de quadrados seqüenciais equivalentes às somas de quadrados do tipo I fornecidas pelo SAS-GLM. O STATISTICA calcula as somas de quadrados dos efeitos principais [A, B, C(B)] e da interação A*C(B), testando as hipóteses $H_0^{(10)}$, $H_0^{(14)}$, $H_0^{(18)}$ e $H_0^{(20)***}$, mas sem eliminar as caselas que estão ocupadas como o SAS-GLM. Já o BMDP fornece as seguintes somas de quadrados:

Tabela 27 - Análise da variância fornecida pelo BMDP

BMDP4V--UNIVARIATE AND MULTIVARIATE ANALYSIS OF VARIANCE AND COVARIANCE, INCLUDING REPEATED MEASURES
Version: PC90 (1990 IBM PC/MS-DOS) Date: 01/21/98 at 08:53:45

BETWEEN = EQUAL

EFFECT	VARIATE	H ₀	STATISTIC	F	D.F	P
A	SS = 394,3394	H ₀ ⁽¹⁰⁾	MS = 394,3394	1,08	1; 17	0,3125
B	SS = 3623,7643	H ₀ ⁽¹⁴⁾	MS = 3623,7643	9,96	1; 17	0,0058
AB	SS = 478,7826	H ₀ ⁽¹⁶⁾	MS = 478,7826	1,32	1; 17	0,2673
C	SS = 7604,7262	H ₀ ⁽¹⁸⁾	MS = 2534,9087	6,96	3; 17	0,0029
AC	SS = 418,9709	H ₀ ^{(20)***}	MS = 209,4854	0,58	2; 17	0,5730
ERROR	SS = 6188,0000	-	MS = 364,0000			
A B,AB,C	SS = 13,4711	-	MS = 13,4711	0,04	1; 17	0,8497
B A,AB,C	SS = 3340,3545	-	MS = 3340,3545	9,18	1; 17	0,0076
AB A,B,C	SS = 112,3275	H ₀ ^{(16)*}	MS = 112,3275	0,31	1; 17	0,5858
C A,B,AB	SS = 7198,9535	H ₀ ^{(18)**}	MS = 2399,6512	6,59	3; 17	0,0037

BETWENN = SIZES

EFFECT	VARIATE	H ₀	STATISTIC	F	D.F	P
A	SS = 275,0417	H ₀ ⁽⁷⁾	MS = 275,0417	0,76	1; 17	0,3968
B	SS = 4146,2607	H ₀ ⁽¹³⁾	MS = 4126,2607	11,39	1; 17	0,0036
AB	SS = 880,4265	H ₀ ⁽¹⁵⁾	MS = 880,4265	2,42	1; 17	0,1383
C	SS = 7810,3611	H ₀ ⁽¹⁹⁾	MS = 2603,4537	7,15	3; 17	0,0026
AC	SS = 418,9709	H ₀ ^{(20)***}	MS = 209,4854	0,58	2; 17	0,5730
A B,AB,C	SS = 0,3176	-	MS = 0,3176	0,00	1; 17	0,9768
B A,AB,C	SS = 3903,7079	-	MS = 3903,7079	10,72	1; 17	0,0045
AB A,B,C	SS = 189,6731	H ₀ ^(***)	MS = 189,6731	0,52	1; 17	0,4802
C A,B,AB	SS = 7116,2214	H ₀ ^{(17)**}	MS = 2372,0738	6,52	3; 17	0,0039

* - Hipótese sem considerar a média μ_{122}

** - Hipóteses sem considerar a média μ_{221}

*** - Hipótese sem considerar as médias μ_{121} , μ_{122} e μ_{221}

H₀^(***) - Hipótese complexa equivalente àquela testada pelo SAS-GLM

O comando BETWEEN = EQUAL do BMDP fornece dois quadros para a

análise da variância. No primeiro quadro, as somas de quadrados dos efeitos principais, A, B, C(B), e da interação A*B testam as hipóteses do tipo IV sobre as médias não ponderadas, sem eliminar as caselas que estão ocupadas. Nesse caso, suas hipóteses são diferentes daquelas testadas pelo SAS-GLM, pois conforme a posição e o número de caselas vazias, o SAS-GLM elimina uma ou mais caselas que estão ocupadas. Já a soma de quadrados da interação A*C(B) fornecida pelo BMDP testa hipótese equivalente àquela testada pelo SAS-GLM. No segundo quadro, as somas de quadrados dos efeitos principais, A e B, fornecidas pelo BMDP testam hipóteses desconhecidas, porém, as somas de quadrados da interação A*B e do fator C(B) testam as hipóteses do tipo III sobre as médias não ponderadas equivalentes àquelas testadas pelo SAS-GLM, conforme consta na Tabela 19.

O comando BETWEEN = SIZES do BMDP também fornece dois quadros para a análise da variância, onde no primeiro quadro, suas somas de quadrados testam as hipóteses do tipo I sobre as médias ponderadas não ajustadas e, portanto são diferentes daquelas testadas pelo SAS-GLM. No segundo quadro, as somas de quadrados dos efeitos principais A e B testam hipóteses desconhecidas. Já as somas de quadrados da interação A*B e do fator C(B) são equivalentes às somas de quadrados do tipo II fornecidas pelo SAS-GLM. Testam, portanto, as hipóteses do tipo II, no caso da interação A*B, gerada a partir de funções estimáveis complexas do tipo II, e no caso do fator C(B), sobre as médias ponderadas ajustadas, conforme consta na Tabela 19.

A seguir, com base nos exemplos 1 e 2, são apresentadas um resumo dos resultados obtidos em termos da performance dos *softwares* estatísticos utilizados

Tabela 28 - Performance dos *softwares* estatísticos

Hipóteses Associadas	B M D P	G L I M	M I N I T A B	N T I A	S A E G	S A N E S T	S A S	S T A T G R A P H I C S	S T A T I S T I C A
$H_o^{(1-A)}$	X	X	X	X	X		X		
$H_o^{(2-A)}$	X								X
$H_o^{(3-A)}$							X		
$H_o^{(4-B(A))}$	X	X	X	X	X		X		
$H_o^{(5-B(A))}$	X			X			X		X
$H_o^{(6-C(AB))}$	X	X	X	X	X		X		X
$H_o^{(7-A)}$	X	X	X	X			X		
$H_o^{(8-A)}$		X*	X*	X*			X*		
$H_o^{(9-A)}$							X		
$H_o^{(10-A)}$	X								X
$H_o^{(11-A)}$							X		
$H_o^{(12-B)}$		X	X	X			X		
$H_o^{(13-B)}$	X	X*	X*	X*			X*		
$H_o^{(14-B)}$	X			X			X		X
$H_o^{(15-AB)}$	X	X	X	X			X		
$H_o^{(16-AB)}$	X			X			X		
$H_o^{(17-C(B))}$	X	X	X	X			X		
$H_o^{(18-C(B))}$	X			X			X		X
$H_o^{(19-C(B))}$	X	X**	X**	X**			X**		
$H_o^{(20-AC(B))}$	X	X	X	X			X		X

* - Se considerada a ordem [B-A-AB-C(B)-AC(B)]

** - Se considerada a ordem [B-C(B)-A-AB-AC(B)]

5 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 os resultados 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 e na interpretação dos resultados.

Face aos resultados obtidos, concluiu-se que:

1 - Modelo Com Três Fatores Hierarquizados

1.a - Hipóteses testadas pelo procedimento GLM do SAS

- 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. Testam, portanto, as hipóteses $H_o^{(1)}$, $H_o^{(4)}$ e $H_o^{(6)}$, respectivamente.

- A soma de quadrados do tipo II referente ao fator A, fornecida pelo SAS-GLM, não testa a hipótese do tipo II sobre as médias ponderadas ajustadas para o fator B, como nos casos de esquemas fatoriais, pois o fator B está aninhado sob o fator A. Nesse caso, a hipótese associada à soma de quadrados do tipo II é equivalente à hipótese do tipo I, $H_o^{(1)}$.

- Quando o número de níveis do fator $C(A\ B)$ são diferentes, independente 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 sobre as médias não ponderadas, $H_o^{(2)}$, mas uma hipótese não usual, baseada em funções estimáveis complexas do tipo III, $H_o^{(3)}$.

1.b - Hipóteses testadas por outros *softwares* estatísticos

- Dos *softwares* estudados, apenas o SANEST e o STATGRAPHICS não realizam as análises para 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 e, portanto, testam hipóteses equivalentes.

- 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 parcial referente

ao fator A fornecida pelo NTIA, 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 não realiza as análises para modelos hierarquizados com dados desbalanceados, mas 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. Nesse caso, testam as hipóteses $H_0^{(2)}$, $H_0^{(5)}$ e $H_0^{(6)}$, respectivamente.

- O comando BETWEEN = EQUAL do BMDP fornece somas de quadrados equivalentes às somas de quadrados do tipo III do SAS-GLM. Se existem caselas vazias, então, as somas de quadrados obtidas pelo comando BETWEEN = EQUAL do BMDP são 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 do tipo I fornecidas pelo SAS-GLM.

2 - Modelo Com Estrutura do Tipo Fatorial-Hierárquica

2.a - Hipóteses testadas pelo procedimento GLM do SAS.

- Das somas de quadrados do tipo I referentes aos efeitos principais A, B e C(B) fornecidas pelo SAS-GLM, apenas a soma de quadrados do fator A testa a hipótese do tipo I sobre as médias ponderadas não ajustadas, $H_0^{(7)}$. As demais somas de quadrados do tipo I, fornecidas pelo SAS-GLM testam as hipóteses, $H_0^{(12)}$ e $H_0^{(17)}$, que não corres-

pondem às hipóteses do tipo I, $H_o^{(13)}$ e $H_o^{(19)}$.

- A soma de quadrados do tipo II referente ao fator A, $R[\alpha | \mu, \beta, \gamma(\beta)]$, fornecida pelo SAS-GLM, testa a hipótese do tipo II sobre as médias ponderadas ajustadas para os fatores B e C(B), $H_o^{(9)}$, porém, a soma de quadrados do tipo II referente ao fator B, $R(\beta | \mu, \alpha)$, testa a hipótese do tipo II sobre as médias ponderadas ajustadas somente para o fator A, $H_o^{(12)}$, pois o fator C(B) está aninhado sob o fator B.

- A soma de quadrados do tipo II referente a interação A*B fornecida pelo SAS-GLM testa hipótese do tipo II, mas são complexas e de difícil interpretação.

- Quando os níveis do fator C(B) são diferentes, independentemente dos dados serem balanceados ou não, a soma de quadrados do tipo III fornecida pelo SAS-GLM, também não testa a hipótese sobre as médias não ponderadas, $H_o^{(10)}$, mas uma hipótese não usual baseada na soma de quadrados gerada a partir de funções estimáveis complexas do tipo III, $H_o^{(11)}$.

2.b - Hipóteses testadas por outros *softwares* estatísticos

- Dos *softwares* estudados, o SANEST, o SAEG e o STATGRAPHICS não realizam as análises para modelos com estrutura do tipo fatorial-hierárquica.

- O MINITAB, o GLIM e o NTIA fornecem somas de quadrados do tipo sequencial, equivalentes às somas de quadrados do tipo I do SAS-GLM.

- O NTIA também fornece as somas de quadrados parciais, equivalentes às somas de quadrados do tipo III do SAS-GLM, exceto para a soma de quadrados do fator A,

que não testa nem a hipótese $H_o^{(11)}$ e nem a hipótese sobre as médias não ponderadas, $H_o^{(10)}$.

- O STATISTICA não realiza as análises para modelos com dados desbalanceados, mas é possível obter as somas de quadrados dos efeitos principais A, B, C(B) e da interação A*C(B), equivalentes às somas de quadrados do tipo IV fornecidas pelo SAS-GLM, através do comando “CONTRASTS FOR BETWEEN-GROUP FACTORS”.

- O comando BETWEEN = EQUAL do BMDP fornece somas de quadrados equivalentes às somas de quadrados do tipo III do SAS-GLM. Se existem caselas vazias, então, o comando BETWEEN = EQUAL do BMDP fornece somas de quadrados do tipo IV, levando em conta todas as caselas ocupadas.

- O comando BETWEEN = SIZES do BMDP fornece somas de quadrados do tipo I e, ao contrário do SAS-GLM, testa hipóteses do tipo I sobre as médias ponderadas não ajustadas.

REFERÊNCIAS BIBLIOGRÁFICAS

- ANDERSON, V. L.; MCLEAN, R. A. **Design of experiments - A realistic approach.** New York: Marcel Dekker Inc., 1974, 471p.
- CAMARINHA FILHO, J. A. Testes de hipóteses em modelos lineares com dados desbalanceados e caselas vazias, Piracicaba, 1995, 142p. Dissertação (Mestrado em Estatística e Experimentação Agronômica), ESALQ-USP.
- CARLSON, J. E.; TIMM, N. H. Analysis of nonorthogonal fixed-effects designs. **Psychological Bulletin**, v.81, p.563-570, 1974.
- COMSTOCK, R. E. Overestimation of mean squares by the method of expected numbers. **Journal of the American Statistical Association**, v.38, p.335-340, 1943.
- CRAMER, E. M. A nonorthogonal analysis of variance program. **Journal of the American Statistical Association**, v.71, p.93-95, 1976.
- 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. **Commun. Statist. - Simula.**, v.15, p.215-225, 1986.
- ELSTON, R. C.; BUSH, N. The hypotheses that can be tested when there are interactions in an analysis of variance model. **Biometrics**, v.20, p.681-698, 1964.

- FEDERER, W. T.; ZELEN, M. Analysis of multifactor classifications with unequal numbers of observations. **Biometrics**, v.22, p.525-552, 1966.
- FINNEY, D. J. Main effects and interactions. **Journal of the American Statistical Association**, v.43, p.566-571, 1948.
- FRANCIS, I. A comparison of several analysis of variance programs. **Journal of the American Statistical Association**, v.68, p.860-865, 1973.
- GOODNIGHT, J. H. General linear model procedure. **SAS.ONE, Proceedings of the First International Users Conference**, Kissimee, Florida, p.1-39, 1976.
- GOODNIGHT, J. H. Test of hypothesis in fixed effect linear models. **Commun. Statist. - Theor. Meth.**, A9(2), p.167-180, 1980.
- GOSSLEE, D. G.; LUCAS, H. L. Analysis of variance of disproportionate data when interaction is present. **Biometrics**, v.21, p.115-133, 1965.
- HAMER, R. M. Papers that evaluate computer programs. **The American Statistician**, v.35, p.264, 1981.
- HARVEY, W. R. **Least squares analysis of data with unequal subclass numbers**. U. S. Dept. of Agriculture Publication, ARS 20-8, 1960.
- HARVEY, W. R. **Least squares analysis of data with unequal subclass numbers**. U. S. Dept. of Agriculture Publication, ARS 20-8, July, 1968.
- HAZEL, L. N. The covariance analysis of multiple classification tables with unequal subclass numbers. **Biometrics**, v.2, p.21-25, 1946.
- HENDERSON, C. R. Estimation of variance and covariance components. **Biometrics**, v.9, p.226-252, 1953.

- HENDRICKS, W. A. Analysis of variance considered as an application of simple error theory. **Annals of Mathematical Statistics**, v.6, p.117-126, 1934.
- HERR, D. G. On the history of anova in unbalanced, factorial 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.
- HOCKING, R. R.; SPEED, F. M. A full rank analysis of some linear model problems. **Journal of the American Statistical Association**, v.70, p.706-712, 1975.
- HOCKING, R. R.; SPEED, F. M.; COLEMAN, A. T. Hypotheses tested with unbalanced data. **Commun. Statist. - Theor. Meth.**, A9(2), p.116-129, 1980.
- 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**. Colombia: 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. **4º Congresso Brasileiro de Usuários do SAS**. Universidade de São Paulo. 1995-b, 111p.
- IEMMA, A. F. Dados estatísticos desbalanceados: PROC SAS/GLM. **II Semana de Matemática**. Universidade Federal Rural do Rio de Janeiro, 1995-c, 130p.
- IEMMA, A. F. **Análisis de varianza de experimentos con celdas vazias**. Trabajos de

Matematicas – Serie C – N. 22/97. Universidad Nacional de Cordoba, Argentina. 1997, 112p.

IEMMA, A. F.; PERRI, S. H. V. **Ajuste de modelos mistos através do sistema estatístico SAS**. DME – ESALQ/USP. 1996, 99p.

KRAMER, C. Y. On the analysis of variance of a two-way classification with unequal subclass numbers. **Biometrics**, v.11, p.441-452, 1955.

LINDMAN, H. R. **Analysis of variance in complex experimental designs**, San Francisco: W. H. Freeman and Company, 1974, 352p.

MONDARDO, M. Estimabilidade de funções paramétricas com dados desbalanceados através do PROC-GLM do SAS: Aplicações à pesquisa agropecuária. Piracicaba, 1944, 166p. Dissertação (Mestrado em Estatística e Experimentação Agronômica), ESALQ-USP.

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.; KLETT, J. C. **Applied multivariate analysis**. New York: McGraw-Hill, 1972.

OVERALL, J. E.; SPIEGEL, D. K. Concerning least squares analysis of experimental data. **Psychological 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 em Estatística e Experimentação Agronômica), ESALQ-USP.

- PATTERSON, R. E. The use of adjusting factors in analysis of data with disproportionate subclass numbers. **Journal of the American Statistical Association**, v.41, p.334-346, 1946.
- SANTOS, E. S. Utilização de "*Softwares*" estatísticos na interpretação de hipóteses com dados desbalanceados. Piracicaba, 1994, 175p. Tese (Doutorado em Estatística e Experimentação Agronômica), ESALQ-USP.
- SCHEFFÉ, H. **The analysis of variance**, New York: John Wiley & Sons, 1959, 477p.
- SEARLE, S. R. Additional results concerning estimable functions and generalized inverse matrices. **J. Roy. Statist. Soc., Série B**, v.27, p.486-490, 1965.
- SEARLE, S. R. **Linear models**. New York: John Wiley & Sons, 1971, 532p.
- SEARLE, S. R. Arbitrary hypotheses in linear models with unbalanced data. **Commun. Statist. - Theor. Meth.**, A9(2), p.181-200, 1980.
- SEARLE, S. R.; SPEED, F. M.; HENDERSON, H. V. Some computational and model equivalences in analysis of variance unequal-subclass-numbers data. **The American Statistician**, v.35, p.16-33, 1981.
- SEARLE, S.R. **Linear models for unbalanced data**, New York: John Wiley & Sons, 1987, 536p.
- SEARLE, S. R. Analysis of variance computing package output for unbalanced data from fixed-effects models with nested factors. **The American Statistical Association**, v.48, p.148-153, 1994.

- SNEDECOR, G. W. The method of expected numbers for tables of multiple classification with disproportionate subclass numbers. **Journal of the American Statistical Association**, v.29, p.389-393, 1934.
- SNEDECOR, G. W.; COX, G. Disproportionate subclass numbers in tables of multiple classification. **Iowa state Agriculture Experiment Station Research Bulletin**, v.180, 1935.
- SPEED, F. M.; HOCKING, R. R. The use of the R()-notation with unbalanced data. **The American Statistician**, v.30, p.30-33, 1976.
- SPEED, F. M.; HOCKING, R. R.; HACKNEY, O. P. Methods of linear models with unbalanced data. **Journal of the American Statistical Association**, v.73, p.105-112, 1978.
- URQUHART, N. S.; WEEKS, D. L. Linear models in messy data: some problems and alternatives. **Biometrics**, v.34, p.896-705, 1978.
- VAJDA, S. Technique of the analysis of variance. **Nature**, v.160, p.27, 1947.
- WINER, B. J. **Statistical principles in experimental design**, 2nd edition, New York: McGraw- Hill Book Co, 1971, 907p.
- YATES, F. The principles of orthogonality and confounding in replicated experiments. **J. Agric. 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.
- YATES, F. Technique of the analysis of variance. **Nature**, v.160, p.472-473, 1947.