ANÁLISE MULTIVARIADA APLICADA AS CIÊNCIAS AGRÁRIAS

PÓS-GRADUAÇÃO EM AGRONOMIA CIÊNCIA DO SOLO: CPGA-CS

ANÁLISE DE VARIÂNCIA MULTIVARIADA

Carlos Alberto Alves Varella[1]

ÍNDICE

[1] Professor. Universidade Federal Rural do Rio de Janeiro, IT-Departamento de Engenharia, BR 465 km 7 - CEP 23890-000 – Seropédica – RJ. E-mail: varella@ufrrj.br.

INTRODUÇÃO

A análise de variância multivariada é utilizada para comparar vetores de médias. Os dados normalmente são provenientes de delineamentos estatísticos. A formulação de um teste estatístico para comparar vetores de médias, depende da partição do total da variância em: variância devido ao efeito de tratamentos e variância devido ao erro. Está partição da variância total é denominada de MANOVA, análise de variância multivariada (JOHNSON & WICHERN, 1999). Em experimentos que envolvem variáveis aleatórias contínuas, medidas na mesma unidade experimental, pode-se pressupor a multinormalidade e realizar uma análise multivariada. Um ponto relevante da análise multivariada é o aproveitamento da informação conjunta das variáveis envolvidas (REGAZZI, 2000).

As pressuposições para realização da MANOVA são as seguintes:

1) Modelo aditivo para efeitos de tratamentos, blocos (se houver) e erro;
2) Independência dos erros;
3) Igualdade da matriz de covariância Σ para todas as amostras;
4) Distribuição multinormal dos erros, com variância Σ.

MODELO ESTATÍSTICO

O modelo estatístico para um delineamento em blocos casualizados com b blocos e k tratamentos em que são medidas p variáveis é:

$$Y_{ijr} = \mu_r + t_{ir} + b_{jr} + e_{ijr}$$

$i = 1,2,\ldots,k; \quad j = 1,2,\ldots,b; \quad r = 1,2,\ldots,p;$

em que,

r	=	indexador das variáveis;
Y_{ijr}	=	valor observado da r-ésima variável sob o efeito do i-ésimo tratamento no j-ésimo bloco;
μ_r	=	média geral da r-ésima variável;
t_{ir}	=	efeito do i-ésimo tratamento na r-ésima variável;
b_{jr}	=	efeito do j-ésimo bloco na r-ésima variável;
e_{ijr}	=	efeito aleatório associado à observação Y_{ijr}

Na forma matricial o modelo estatístco é:

$$Y = X\mathrm{B} + \varepsilon$$

em que,

Y = matriz de observações de dimensões *bk* x *p;*
X = matriz do delineamento de dimensões *bk* x *(1+k+b)*;
B = matriz de parâmetros de dimensões *(1+k+b)* x *p*;
ε = matriz de erros de dimensões *bk* x *p*.

Podemos também escrever da seguinte forma:

$$Y = \left[\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_p\right] = \left[X\tilde{\beta}_1, X\tilde{\beta}_2, \ldots, X\tilde{\beta}_p\right] + \left[\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_p\right]$$

Para a variável *r (r=1,2,...,p)* temos que:

$$\tilde{y}_r = X\tilde{\beta}_r + \tilde{e}_r$$

MATRIZES DE SISTEMAS DE EQUAÇÕES NORMAIS

Considerando-se o modelo linear multivariado $Y = X\mathrm{B} + \varepsilon$, a matriz de sistemas de equações normais é:

$$X'X\dot{\mathrm{B}} = X'Y$$

A solução de mínimos quadrados do sistema de equações normais é dada pela matriz:

$$\dot{\mathrm{B}} = \left[\dot{\beta}_1, \dot{\beta}_2, \ldots, \dot{\beta}_p\right] = \left(X'X\right)^{-1} X'Y$$

A matriz de soluções $\dot{\mathrm{B}}$ possui dimensões *(1+k+b)* x *p*, onde X é a matriz de valores das p variáveis independentes e Y é a matriz de valores resposta para os efeitos de tratamentos. Da mesma forma que no modelo univariado, sendo posto de (X’X)=k+b-1, o sistema admite mais de uma solução, isto é, a matriz (X’X) não possui inversa verdadeira. Impondo-se as restrições de que os somatórios dos efeitos de tratamentos e blocos sejam igual a zero, o sistema apresenta solução única que denotaremos por:

$$\hat{\mathrm{B}} = \left(X'X + AA'\right)^{-1} X'Y = \left[\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right]$$

em que,

$\hat{\mathrm{B}}$ = melhor estimador linear ou BLUE de B;
$\left(X'X + AA'\right)^{-1}$ = inversa condicional de X’X.

DECOMPOSIÇÃO DA SOMA DE QUADRADOS E PRODUTOS TOTAIS

Considerando-se o modelo linear multivariado $Y = X\mathrm{B} + \varepsilon$, tem-se, do método dos mínimos quadrados, que:

$$\hat{\varepsilon}'\hat{\varepsilon} = Y'Y - \hat{\mathrm{B}}'X'Y$$

em que,

$\hat{\varepsilon}'\hat{\varepsilon}$ = matriz de soma de quadrados e produtos de resíduo que denotaremos por E, com n_e= n-Posto[X] graus de liberdade. No modelo de blocos ao acaso temos que:

$$n_e = kb - (k + b - 1) = kb - k - b + 1 = (k-1)(b-1)$$

$Y'Y$ = matriz de soma de quadrados e produtos de totais que denotaremos por A;

$\hat{\mathrm{B}}'X'Y$ = matriz de soma de quadrados e produtos de parâmetros;

Da mesma forma que no modelo univariado temos que:

$$A = H + B + E$$

onde,

$A, H, B \; e \; E$ são matrizes de dimensões p x p, de soma de quadrados e produtos de totais, tratamentos, blocos e resíduo, respectivamente.

TESTE PARA A HIPÓTESE NULA H_0

A hipótese H_0 a ser testada, considerando k tratamentos e p variáveis, é a de que os vetores de médias de tratamentos são iguais, isto é:

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_k$$

ou

$$H_0 : \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \cdots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \cdots \\ \mu_{2p} \end{bmatrix} = \cdots = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \\ \cdots \\ \mu_{kp} \end{bmatrix}$$

Essa hipótese é o mesmo que testar se os vetores de efeitos de tratamentos são nulos, isto é:

$$H_0 : \tilde{t}_1 = \tilde{t}_2 = \cdots = \tilde{t}_k = 0$$

ou

$$H_0: \begin{bmatrix} t_{11} \\ t_{12} \\ \cdots \\ t_{1p} \end{bmatrix} = \begin{bmatrix} t_{21} \\ t_{22} \\ \cdots \\ t_{2p} \end{bmatrix} = \cdots = \begin{bmatrix} t_{k1} \\ t_{k2} \\ \cdots \\ t_{kp} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

O teste de Wilks é o mais utilizado para testar a hipótese H_0 da MANOVA. Contudo outros testes também são utilizados, tais como Pillai, Hotelling-Lawley e o teste de Roy, os quais podem apresentar resultados diferentes para a mesma análise. O teste de Wilks é representado pela letra grega Λ (lambda maiúsculo), assim definido:

$$\Lambda = \frac{\det(E)}{\det(H+E)} = \frac{|E|}{|H+E|}$$

Na presença de diferenças sistemáticas entre tratamentos, espera-se sempre obter $\Lambda < 1$, e tanto mais significativo quanto menor for seu valor. Para avaliar a significância do valor de Λ obtido, pode-se usar tabela própria para o teste de Wilks, ou trnsformar o valor de Λ em valor F correspondente. O valor de Λ obtido na tabela para o teste de Wilks é função de α, p, q e n_e. O valaor de Λ tabelado é:

$$\Lambda_{tab} = \Lambda(\alpha; p; q; n_e)$$

A regra de decisão é: rejeita-se a hipótese nula ao nível de significância α se:

$$\Lambda_{cal} < \Lambda_{tab}$$

Caso contrário não se rejeita H_0 e diz-se que o teste foi significativo ao nível de siginificância α.

TESTE PARA VETORES DE MÉDIAS

Quando rejeitamos a hipótese H_0, isto é, os vetores de médias não são iguais, e temos mais de dois tratamentos, torna-se necessário um teste de médias para verificar quais vetores são diferentes entre si. Para comparar dois vetores de médias de tratamentos, podemos usar o teste T^2 de Hotelling como apresentado a seguir:

EXEMPLO DE APLICAÇÃO

Fazer MANOVA para experimento com um fator inteiramente casualizado utilizando o programa computacional SAS.

BIBLIOGRAFIA

FISHER, R.A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v.**7**, p.179-188, 1936.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999, 815 p.

KHATTREE, R. & NAIK, D.N**. Multivariate data reduction and discrimination with SAS software**. Cary, NC, USA: SAS Institute Inc., 2000. 558 p.

REGAZZI, A.J. Análise multivariada, notas de aula INF 766, Departamento de Informática da Universidade Federal de Viçosa, v.2, 2000.

VARELLA, C.A.A. **Estimativa da produtividade e do estresse nutricional da cultura do milho usando imagens** digitais. 2004. 92 f. Tese (Doutorado em Engenharia Agrícola) – Universidade Federal de Viçosa, Viçosa, 2004.

SAS. Online doc version 8. Disponível em: http://v8doc.sas.com/sashtml/. Acesso em 14 mar. 2007.