6 Análise de variância multivariada

6.1. Introdução

Com o desenvolvimento da estatística no século XX a possibilidade de condução e análise de experimentos propiciou grande sucesso às pesquisas, principalmente pela habilidade de lidar com variações não controláveis. O primeiro a representar os resultados experimentais por um modelo foi W. S. Gosset (Student, 1908).

As terminologias dos delineamentos experimentais, independentemente da área de aplicação, tenderam a serem iguais aos dos experimentos em agricultura. Portanto, unidades experimentais são denominadas de parcelas e o valor da variável aleatória como resposta. Experimentos com apenas uma classificação dos tratamentos são denominados de delineamentos inteiramente casualizados ou de classificação simples. Experimentos em que vários tipos de tratamentos são aplicados ao material experimental simultaneamente são denominados de fatoriais. Outra classe de experimentos é gerada pelos arranjos hierarquizados dos materiais.

O presente capítulo tem por objetivo apresentar a extensão multivariada dos métodos univariados de análise de variância. As idéias básicas desse capítulo podem ser estendidas a todos os tipos de delineamentos e arranjos das estruturas de tratamentos, embora sejam apresentas na situação mais simples, qual seja, delineamento de classificação simples.

6.2. Delineamento de classificação simples

O caso mais simples dos delineamentos experimentais é o de classificação simples ou delineamento inteiramente casualizado. O arranjo experimental consiste em g tratamentos, possivelmente incluindo a(s) testemunha(s), para os quais as unidades experimentais são aleatorizadas.

As amostras aleatórias de cada tratamento são representadas por:

Tratamento 1: $\underset{\sim}{X}_{11}$, $\underset{\sim}{X}_{12}$, ..., $\underset{\sim}{X}_{1n_1}$

Tratamento 2: $\underset{\sim}{X}_{21}$, $\underset{\sim}{X}_{22}$, ..., $\underset{\sim}{X}_{2n_2}$

$\vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots$

Tratamento g: $\underset{\sim}{X}_{g1}$, $\underset{\sim}{X}_{g2}$, ..., $\underset{\sim}{X}_{gn_g}$

A análise de variância multivariada (MANAVA) é usada para investigar se os vetores de médias de tratamento são os mesmos, e se não, qual componente de média difere significativamente. Algumas pressuposições da estrutura dos dados devem ser obedecidas para validade da inferência estatística: (a) $\underset{\sim}{X}_{\ell 1}, \underset{\sim}{X}_{\ell 2}, \cdots, \underset{\sim}{X}_{\ell n_\ell}$ deve ser uma amostra aleatória de tamanho n_ℓ do tratamento ℓ, com média $\underset{\sim}{\mu}_\ell$, ℓ=1, 2, ..., g. As amostras dos tratamentos devem ser independentes; (b) todos os tratamentos possuem covariância comum Σ; e (c) cada tratamento tem distribuição normal multivariada.

O modelo de análise de variância multivariada está apresentado a seguir. Neste modelo cada componente é um vetor de p componentes.

$$\underset{\sim}{X}_{\ell j} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_\ell + \underset{\sim}{e}_{\ell j} \qquad j = 1, 2, \cdots, n_\ell \text{ e } \ell = 1, 2, \cdots, g \tag{6.1}$$

em que, $\underset{\sim}{e}_{\ell j}$ é independentemente e identicamente distribuído e $N_p(0, \Sigma)$ para todo ℓ e j; $\underset{\sim}{\mu}$ é o vetor média geral e $\underset{\sim}{\tau}_\ell$ representa o vetor de efeitos do ℓ-ésimo tratamento. Pode-se adotar a restrição paramétrica $\sum_{\ell=1}^{g} n_\ell \underset{\sim}{\tau}_\ell = \underset{\sim}{0}$.

Os erros do vetor $\underset{\sim}{X}_{\ell j}$ são correlacionados, no entanto a matriz de covariância Σ é a mesma para todos os tratamentos.

O vetor de observações pode ser decomposto em:

$$\underset{\sim}{X}_{\ell j} = \overline{\underset{\sim}{X}} + (\overline{\underset{\sim}{X}}_{\ell} - \overline{\underset{\sim}{X}}) + (\overline{\underset{\sim}{X}}_{\ell j} - \overline{\underset{\sim}{X}}_{\ell}) \quad (6.2)$$

Observação = Estimativa da média geral + Estimativa do efeito do tratamento + resíduo

Analogamente, demonstra-se que a soma de quadrados e produtos totais possui a seguinte decomposição:

Soma de quadrados e produtos (SQP) total corrigido = SQP tratamentos + SQP resíduo

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\underset{\sim}{X}_{\ell j} - \overline{\underset{\sim}{X}}\right)\left(\underset{\sim}{X}_{\ell j} - \overline{\underset{\sim}{X}}\right)' = \sum_{\ell=1}^{g} n_\ell \left(\overline{\underset{\sim}{X}}_{\ell} - \overline{\underset{\sim}{X}}\right)\left(\overline{\underset{\sim}{X}}_{\ell} - \overline{\underset{\sim}{X}}\right)' + \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\underset{\sim}{X}_{\ell j} - \overline{\underset{\sim}{X}}_{\ell}\right)\left(\underset{\sim}{X}_{\ell j} - \overline{\underset{\sim}{X}}_{\ell}\right)' \quad (6.3)$$

A soma de quadrados e produtos do resíduo pode ser expressa por:

$$E = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\underset{\sim}{X}_{\ell j} - \overline{\underset{\sim}{X}}_{\ell}\right)\left(\underset{\sim}{X}_{\ell j} - \overline{\underset{\sim}{X}}_{\ell}\right)' = (n_1 - 1)S_1 + (n_2 - 1)S_2 + \ldots + (n_g - 1)S_g \quad (6.4)$$

em que S_ℓ é a matriz de covariância amostral do ℓ-ésimo tratamento.

O teste da hipótese de inexistência de efeitos de tratamentos,

$$H_0 : \underset{\sim}{\tau}_1 = \underset{\sim}{\tau}_2 = \cdots = \underset{\sim}{\tau}_g = \underset{\sim}{0} \quad (6.5)$$

é realizado considerando as magnitudes das somas de quadrados e produtos de tratamento e resíduo pela variância generalizada.

O esquema de análise de variância multivariada (MANAVA) está apresentado na Tabela 6.1. A fonte de variação total é particionada em causas de variação devido a tratamento e ao erro experimental ou resíduo.

Tabela 6.1. Tabela de MANAVA para testar a hipótese de igualdade do vetor de efeito dos tratamentos em um delineamento de classificação simples.

FV	GL	Matriz de SQP
Tratamento	g-1	$B = \sum_{\ell=1}^{g} n_\ell \left(\underset{\sim}{\bar{X}}_\ell - \underset{\sim}{\bar{X}}\right)\left(\underset{\sim}{\bar{X}}_\ell - \underset{\sim}{\bar{X}}\right)'$
Resíduo	$\upsilon = \sum_{\ell=1}^{g} n_\ell - g$	$E = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\underset{\sim}{X}_{\ell j} - \underset{\sim}{\bar{X}}_\ell\right)\left(\underset{\sim}{X}_{\ell j} - \underset{\sim}{\bar{X}}_\ell\right)'$
Total corrigido	$\sum_{\ell=1}^{g} n_\ell - 1$	$B + E = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\underset{\sim}{X}_{\ell j} - \underset{\sim}{\bar{X}}\right)\left(\underset{\sim}{X}_{\ell j} - \underset{\sim}{\bar{X}}\right)'$

Os critérios para o teste da hipótese apresentada em (6.5), envolvem variâncias generalizadas e autovalores e autovetores da maximização de duas formas quadráticas dadas em (2.15 e 2.16).

De maneira geral, supondo que H seja a matriz de SQP relativa aos efeitos dos tratamentos que se deseja testar a igualdade, para o exemplo H=B, então a solução da equação determinantal dada por:

$$| H - \lambda_i E | \underset{\sim}{e}_i = 0$$

fornece as estimativas dos autovalores e autovetores, necessários aos testes de hipótese (6.5), os quais estão apresentados na Tabela 6.2. Quatro critérios existem para o teste desta hipótese. Muitos autores recomendam utilizar o critério de Wilks como referência, por se tratar de um teste baseado na razão de verossimilhança. Outros recomendam que a hipótese nula deva ser rejeitada se pelo menos três dos quatro critérios forem significativos em um nível nominal de significância previamente adotado. Esses critérios podem ser aproximados pela distribuição F. Essas aproximações, também, se encontram apresentadas na Tabela 6.2.

Tabela 6.2. Estatísticas multivariadas e suas equivalência aproximada com a distribuição F.

Critério	Estatística	Aproximação F	GL de F
Wilks	$\Lambda=\frac{\lvert E\rvert}{\lvert H+E\rvert}=\prod\frac{1}{1+\lambda_i}$	$F=\left(\frac{1-\Lambda^{\frac{1}{t}}}{\Lambda^{\frac{1}{t}}}\right)\left(\frac{rt-2f}{pq}\right)$	v_1=pq v_2=rt-2f
Traço de Pillai	$V=tr[H(H+E)^{-1}]=\sum\frac{\lambda_i}{1+\lambda_i}$	$F=\left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right)$	v_1=s(2m+s+1) v_2=s(2n+s+1)
Traço de Hotelling Lawley	$U=tr(HE^{-1})=\sum\lambda_i$	$F=\frac{2(sn+1)U}{s^2(2m+s+1)}$	v_1=s(2m+s+1) v_2=2(sn+1)
Raíz máxima de Roy	$\theta=\lambda_1$	$F=\frac{\theta(\nu-d+q)}{d}$	v_1=d $v_2=\nu-d+q$
	p: número de variáveis = posto(H+E); q: GL de tratamento (ou do contraste); ν: GL do erro; S=min(p,q); r=ν- (p-q+1)/2; f=(pq-2)/4; d=max(p,q); m=(|p-q|-1)/2; n=(ν-p-1)/2; e $t=\begin{cases}\sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & \text{Se } p^2+q^2-5>0\\ 1 & \text{cc}\end{cases}$ Obs. Critério de Wilks possui aproximação exata de F se min(p,q)≤2		

Exemplo 6.1

Num experimento envolvendo 4 variedades de feijão, avaliou-se na seca, a produtividade (P) em kg/ha e número de grão por vagem (NGV), utilizando 5 repetições. Os resultados obtidos foram:

	Cultivar							
	A		B		C		D	
P	NGV	P	NGV	P	NGV	P	NGV	
1082	4,66	1163	5,52	1544	5,18	1644	5,45	
1070	4,50	1100	5,30	1500	5,10	1600	5,18	
1180	4,30	1200	5,42	1550	5,20	1680	5,18	
1050	4,70	1190	5,62	1600	5,30	1700	5,40	
1080	4,60	1170	5,70	1540	5,12	1704	5,50	
5462	22,76	5823	27,56	7734	25,90	8328	26,71	

Teste a hipótese de igualdade do vetor média de tratamentos.

Os vetores médias amostrais de tratamento são:

$$\underset{\sim}{\bar{X}}_1 = \begin{bmatrix} 1092,400 \\ 4,552 \end{bmatrix} \quad \underset{\sim}{\bar{X}}_2 = \begin{bmatrix} 1164,600 \\ 5,512 \end{bmatrix} \quad \underset{\sim}{\bar{X}}_3 = \begin{bmatrix} 1546,800 \\ 5,180 \end{bmatrix} \quad \underset{\sim}{\bar{X}}_4 = \begin{bmatrix} 1665,600 \\ 5,342 \end{bmatrix}$$

E a média geral:

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} 1367,35000 \\ 5,1465 \end{bmatrix}$$

A matriz B é obtida por:

$$B = 5\left\{\begin{bmatrix}1092,400\\ 4,552\end{bmatrix} - \begin{bmatrix}1367,3500\\ 5,512\end{bmatrix}\right\}\left\{\begin{bmatrix}1092,400 & 4,552\end{bmatrix} - \begin{bmatrix}1367,3500 & 5,1465\end{bmatrix}\right\} + \cdots +$$
$$+ 5\left\{\begin{bmatrix}1665,600\\ 5,342\end{bmatrix} - \begin{bmatrix}1367,3500\\ 5,512\end{bmatrix}\right\}\left\{\begin{bmatrix}1665,600 & 5,512\end{bmatrix} - \begin{bmatrix}1367,3500 & 5,1465\end{bmatrix}\right\}$$

Obviamente, quando os cálculos não são realizados no computador, é mais fácil de se obter as matrizes de somas de quadrados e produtos, pelas expressões apresentadas a seguir. Para isso, considere que $X_{i\,\ell\,j}$ representa o valor observado na i-ésima variável, do ℓ-ésimo tratamento na j-ésima unidade experimental. Então,

$$SQB_{ii} = \sum_{\ell=1}^{g} \frac{X_{i\ell.}^2}{n_\ell} - \frac{X_{i..}^2}{\sum_{\ell=1}^{g} n_\ell} \qquad (6.6)$$

representa a soma de quadrados de tratamento para o i-ésimo componente, e

$$SPB_{ik} = \sum_{\ell=1}^{g} \frac{X_{i\ell.}\ X_{k\ell.}}{n_\ell} - \frac{X_{i..}\ X_{k..}}{\sum_{\ell=1}^{g} n_\ell} \qquad (6.7)$$

representa a soma de produtos de tratamento entre as variáveis i e k, com $i \neq k = 1, 2, ..., p$.

Para o total as SQ e SP são:

$$SQT_{ii} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell} X_{i\ell j}^2 - \frac{X_{i..}^2}{\sum_{\ell=1}^{g} n_\ell} \qquad (6.8)$$

$$SPT_{ik} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell} X_{i\ell j} X_{k\ell j} - \frac{X_{i..} X_{k..}}{\sum_{\ell=1}^{g} n_\ell} \qquad (6.9)$$

Para o resíduo basta obter a diferença:

$$E = T - B \qquad (6.10)$$

No exemplo, as matrizes B, E e T são:

$$B = \begin{bmatrix} 1189302,1500 & 768,3605 \\ 768,3605 & 2,6318 \end{bmatrix}$$

$$T = \begin{bmatrix} 1218360,5500 & 778,2645 \\ 778,2645 & 2,9517 \end{bmatrix}$$

$$E = T - B = \begin{bmatrix} 29058,4000 & 9,9040 \\ 9,9040 & 0,3199 \end{bmatrix}$$

O quadro de MANAVA está apresentado a seguir:

FV	GL	SQ&P
Tratamento	3	$B=\begin{bmatrix} 1189302,1500 & 768,3605 \\ 768,3605 & 2,6318 \end{bmatrix}$
Erro	16	$E=\begin{bmatrix} 29058,4000 & 9,9040 \\ 9,9040 & 0,3199 \end{bmatrix}$
Total Corrigido	19	$T=\begin{bmatrix} 1218360,5500 & 778,2645 \\ 778,2645 & 2,9517 \end{bmatrix}$

Para o teste da hipótese $H_0 : \underset{\sim}{\tau}_1 = \underset{\sim}{\tau}_2 = \cdots = \underset{\sim}{\tau}_g = \underset{\sim}{0}$, a razão entre o par de formas quadráticas $\underset{\sim}{e}_i' B \underset{\sim}{e}_i$ e $\underset{\sim}{e}_i' E \underset{\sim}{e}_i$, deve ser maximizado. Isto eqüivale a resolver o sistema de equação,

$$\left(B - \lambda_i E\right) \underset{\sim}{e}_i = \underset{\sim}{0}$$

Para o exemplo, os autovalores e autovetores são:

$\lambda_1 = 41,3463 \quad \underset{\sim}{e}_1' = \begin{bmatrix} 0,0058 & 0,1952 \end{bmatrix}$

$\lambda_2 = 6,6781 \quad \underset{\sim}{e}_2' = \begin{bmatrix} -0,0012 & 1,7667 \end{bmatrix}$

Alguém desavisado, poderia pensar que o valor do segundo elemento do segundo autovetor (1,7667) fosse algum tipo de erro de digitação, por se tratar de um valor superior a 1. No entanto, isto é perfeitamente possível, pois os autovetores, no caso da maximização da razão entre duas formas

quadráticas, são normalizados da seguinte forma: $\underset{\sim}{e}_i' E \underset{\sim}{e}_i = 1$ e $\underset{\sim}{e}_i' E \underset{\sim}{e}_k = 0 \ (i \neq k)$, o que pode ser facilmente verificado.

Todos os critérios utilizados rejeitaram a hipótese de igualdade dos vetores efeitos tratamento (P<0,01), como pode ser visto no quadro seguinte.

Critério	Estatística	F	G.L.	Pr>F
Wilks	Λ=0,0030756	85,16	v_1=6 e v_2=30	0,0001
Traço de Pillai	V=1,846145	64,00	v_1=6 e v_2=32	0,0001
Traço de Hotelling Lawley	U=48,0244	112,06	v_1=6 e v_2=28	0,0001
Raíz máxima de Roy	θ=41,3463	220,51	v_1=3 e v_2=16	0,0001
	p=2; q=3; v=16; s=2; r=16; f=1; d=3; m=0; n=6,5; e t=2			

6.3. Intervalos de confiança simultâneos para o efeito de tratamentos

Quando a hipótese de efeitos iguais para tratamentos é rejeitada, aqueles efeitos que levaram a rejeição são de interesse. Para comparações simultâneas duas a duas, a aproximação de Bonferroni pode ser usada para construir intervalos de confiança simultâneos para os componentes da diferença $\underset{\sim}{\tau}_k - \underset{\sim}{\tau}_\ell$ (diferenças de efeitos dos tratamentos k e ℓ, respectivamente). Esses intervalos são mais curtos que os obtidos para todos os contrastes, e requerem apenas valores críticos da estatística univariada t.

Fazendo τ_{ki} o i-ésimo componente de $\underset{\sim}{\tau}_k$. Desde que $\underset{\sim}{\tau}_k$ pode ser estimado por $\hat{\underset{\sim}{\tau}}_k = \overline{\underset{\sim}{X}}_k - \overline{\underset{\sim}{X}}$, então,

$$\hat{\tau}_{ki} = \overline{X}_{ki} - \overline{X}_i \tag{6.11}$$

Devido a (6.11) corresponder a diferença entre duas médias amostrais independentes, o teste de t de duas amostras é válido, modificando-se adequadamente o nível de significância. A estimativa da variância do contraste entre duas médias de tratamentos é dada por,

$$\hat{Var}(\overline{X}_{ki} - \overline{X}_{\ell i}) = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{E_{ii}}{GLe} \tag{6.12}$$

A divisão de E_{ii} pelos seus graus de liberdade (GLe), é devido ao fato de que, o elemento em questão (E_{ii}) refere-se a uma soma de quadrados. Desta forma, desde que p variáveis são consideradas e g(g-1)/2 comparações duas a duas serão realizadas, então o intervalo de confiança para diferença de efeitos de tratamento é dado por:

$$\overline{X}_{ki} - \overline{X}_{\ell i} \pm t_{GLe}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{E_{ii}}{GLe}} \tag{6.13}$$

para todos os i = 1, 2, ..., p e todas as diferenças $\ell < k = 1, 2, \ldots, g$.

6.7. Exercício

6.7.1. Repetir a análise de variância do exemplo 6.1 utilizando o “proc GLM” do SAS e solicitar a realização dos seguintes contrastes: i) A e B vs C e D; ii) A vs B e iii) C vs D.

6.8. Referências

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4th edition. Prentice Hall, New Jersey, 1998. 816p.