Modelos lineares normais multivariados

Prof. Caio Azevedo

17 de setembro de 2009

- Suponha um conjunto de G populações independentes da qual retiramos G amostras de tamanho n_i, $i = 1, ..., G$,
- Por suposição, temos que $\mathbf{Y}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, em que i $= 1$, 2,...,G (grupo) e j $= 1,2,...,n_i$ (indivíduo). Notação: Y_{ijk} observação referente à variável k do indivíduo j do grupo i.
- Homocedasticidades: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$.

- Resultando na seguinte matriz de dados ($n = \sum_{i=1}^{G} n_i$):

$$\mathbf{Y}_{(n\times p)} = \begin{bmatrix} Y_{111} & Y_{112} & \cdots & Y_{11p} \\ Y_{121} & Y_{122} & \cdots & Y_{12p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1n_1 1} & Y_{1n_1 2} & \cdots & Y_{1n_1 p} \\ - - - & - - - & - - - & - - - \\ Y_{211} & Y_{212} & \cdots & Y_{21p} \\ Y_{221} & Y_{222} & \cdots & Y_{22p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{2n_2 1} & Y_{2n_2 2} & \cdots & Y_{2n_2 p} \\ - - - & - - - & - - - & - - - \\ \vdots & \vdots & \ddots & \vdots \\ - - - & - - - & - - - & - - - \\ Y_{G11} & Y_{G12} & \cdots & Y_{G1p} \\ Y_{G21} & Y_{G22} & \cdots & Y_{G2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{Gn_G 1} & Y_{Gn_G 2} & \cdots & Y_{Gn_G p} \end{bmatrix}$$

- Comparar $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_G$ vs
  H_1 : pelo menos uma diferença, sob homocedasticidade.
- Uma abordagem: análise de variância multivariada (MANOVA).
- Comparar médias atráves do estudo da decomposição da matrize de covariâncias total.
- Como resumir a informação das matrizes de covariâncias de interesse: variâncias generalizada.

$$\mathbf{Y}_{(n\times p)} = \mathbf{X}_{(n\times q)}\mathbf{B}_{(qxp)} + \boldsymbol{\xi}_{(n\times p)}$$

- $\mathbf{Y}_{(n\times p)}$: matriz de dados
- $\mathbf{X}_{(n\times q)}$: matriz de planejamento (modelo de médias).
- $\mathbf{B}_{(qxp)}$: parâmetros de interesse (médias).
- $\boldsymbol{\xi}_{(n\times p)}$: matriz de resíduos, $\boldsymbol{\xi}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

- Suponha G $= 3$, p $= 2$ e $n_i = 50, i = 1, 2, 3$ (3 grupos, duas variáveis e 50 indivíduos por grupo).
- Modelar as médias: parametrização da casela de referência (grupo 1).
- $\mathbf{Y}_{(150\times 2)}$.
- $$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} \\ \mathbf{1}_{(50\times 1)} & \mathbf{1}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} \\ \mathbf{1}_{(50\times 1)} & \mathbf{0}_{(50\times 1)} & \mathbf{1}_{(50\times 1)} \end{bmatrix}$$
- $$\mathbf{B} = \begin{bmatrix} \mu_1 & \mu_2 \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \end{bmatrix}, \ \alpha_{11} = \alpha_{12} = 0$$
- $\mu_{ik} = \mu_k + \alpha_{ik}$: média da variável k do grupo i.

- Pode-se demonstrar que:

$$\underbrace{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\mathbf{Y}_{ij}-\overline{\mathbf{Y}}\right)\left(\mathbf{Y}_{ij}-\overline{\mathbf{Y}}\right)'}_{\text{Matriz de SQ Total}} = \underbrace{\sum_{i=1}^{G} n_i\left(\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}}\right)\left(\overline{\mathbf{Y}}_i-\overline{\mathbf{Y}}\right)'}_{\text{Matriz de SQ do Modelo}} + \underbrace{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(\mathbf{Y}_{ij}-\overline{\mathbf{Y}}_i\right)\left(\mathbf{Y}_{ij}-\overline{\mathbf{Y}}_i\right)'}_{\text{Matriz de SQ do Resíduo}}$$

$$\mathbf{T} = \mathbf{M}+\mathbf{E}$$

- Seja $\mathbf{\Sigma}_{(p\times p)}$ uma matriz de covariâncias.
- Variância generalizada $|\mathbf{\Sigma}|$ (resume a informação contida em $\mathbf{\Sigma}$).
- Suponha p $= 2$.
- Assim $|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2 - \sigma_{11}^2$.

- Baseado na variância generalizada $\Lambda^* = \frac{|\mathbf{E}|}{|\mathbf{M}+\mathbf{E}|}$.
- (Bartlet) Sob H_0,
  $\Lambda^{**} = -\left(n-1-\frac{p+G}{2}\right)\ln\left(\frac{|\mathbf{E}|}{|\mathbf{M}+\mathbf{E}|}\right) \approx \chi^2_{(p(G-1))}$.
- Quanto menor for Λ^{**} menor será a probabilidade de se rejeitar H_0.
- Nível descritivo $p = P(\Lambda^{**} > \Lambda^{**}_{calc}|\boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_G)$

- Dados da iris: mesmas variáveis anteriormente escolhidas e os três grupos.
- Utilização do pacote *manova* implementado na linguagem R.

  m.dados.iris = iris
  grupos=(m.dados.iris[,5])
  m.Y = cbind(m.dados.iris[,1],m.dados.iris[,2])
  aux1 = diag(1,3,3)
  aux2 = matrix(1,50,1)
  m.X.plan = kronecker(aux1,aux2)
  m.X.plan = m.X.plan[,2:3]
  m.ajuste = manova(m.Y $\sim$ m.X.plan)
- summary.manova(m.ajuste,test="Wilks")
  Wilks (Λ^*) = 0.17, Λ^{**} = 262.20, p -valor < 0.0001.
- Como detectar as diferenças: intervalos de confiança simultâneos. Modelo linear multivariado na forma vetorial.