3.6.1 Modelo lineal general multivariado

La distinción entre los modelos lineales multivariados y los modelos univariados es, como su nombre lo señala, que el modelo multivariado involucra más de una variable dependiente o respuesta.

Considérese que las observaciones multivariadas $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, conforman un conjunto de observaciones independientes de una población normal $p-$variante; es decir, $\boldsymbol{Y}_\alpha \sim N_p(X_\alpha \boldsymbol{\beta}, \boldsymbol{\Sigma})$, para $\alpha = 1, \ldots, n$. Los vectores $\boldsymbol{X}_\alpha$ de tamaño $(1 \times q)$ son conocidos. Tanto la matriz $\boldsymbol{\Sigma}_{p\times p}$, como la matriz $\boldsymbol{\beta}_{q\times p}$ son desconocidas. Los $\boldsymbol{Y}_\alpha$ corresponden a las variables respuesta en un modelo de regresión (dependientes), mientras que las $\boldsymbol{X}_\alpha$ son las variables regresoras o explicativas. En tales condiciones los vectores se pueden relacionar a través de un *modelo lineal general multivariado*, tal como el siguiente:

$$\begin{pmatrix} y_{11} \cdots y_{1p} \\ y_{21} \cdots y_{2p} \\ \vdots \ddots \vdots \\ y_{n1} \cdots y_{np} \end{pmatrix} = \begin{pmatrix} x_{11} \cdots x_{1q} \\ x_{21} \cdots x_{2q} \\ \vdots \ddots \vdots \\ x_{n1} \cdots x_{nq} \end{pmatrix} \begin{pmatrix} \beta_{11} \cdots \beta_{1p} \\ \beta_{21} \cdots \beta_{2p} \\ \vdots \ddots \vdots \\ \beta_{q1} \cdots \beta_{qp} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \cdots \varepsilon_{1p} \\ \varepsilon_{21} \cdots \varepsilon_{2p} \\ \vdots \ddots \vdots \\ \varepsilon_{n1} \cdots \varepsilon_{np} \end{pmatrix}$$

En forma condensada, el modelo lineal multivariado anterior se escribe de la manera siguiente:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{E}$$

La matriz $\boldsymbol{X}$ conforma, en la mayoría de los casos, la matriz de diseño o la matriz de variables regresoras, $\boldsymbol{\beta}$ es la matriz de parámetros desconocidos y la matriz aleatoria $\mathbf{E}$ contiene los errores.

Para los propósitos de este texto, se propone, estima e infiere sobre los modelos ligados una estructura de una y dos vías de clasificación, mediante una conformación adecuada de la matriz de diseño $\boldsymbol{X}$ y de la matriz de parámetros $\boldsymbol{\beta}$. Además, se extiende el análisis de perfiles, de medidas repetidas y de curvas de crecimiento, para el caso de varias poblaciones multivariadas.

Tal como en el modelo lineal clásico ($q = 1$), los estimadores de máxima verosimilitud para $\boldsymbol{\beta}$ y $\boldsymbol{\Sigma}$ son:

$$\widehat{\boldsymbol{\beta}} = \Big(\sum_{\alpha=1}^{n} \boldsymbol{X}'_\alpha \boldsymbol{X}_\alpha\Big)^{-1} \Big(\sum_{\alpha=1}^{n} \boldsymbol{X}'_\alpha \boldsymbol{Y}_\alpha\Big)$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{\alpha=1}^{n} (\boldsymbol{Y}_\alpha - \boldsymbol{X}_\alpha \widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_\alpha - \boldsymbol{X}_\alpha \widehat{\boldsymbol{\beta}})'. \tag{3.36}$$

3.6.3 Análisis de varianza multivariado

Desde un punto de vista práctico, el análisis de varianza multivariado es una técnica con la cual se puede verificar la igualdad de los vectores de medias ligados a varias poblaciones multivariadas.
Muchas hipótesis en el campo multivariado pueden expresarse como las hipótesis concernientes al análisis de regresión esquematizado anteriormente. Dentro de este estilo, se presenta la técnica del análisis de varianza para arreglos de una y dos vías de clasificación.

3.6.4 Modelos de una vía de clasificación

Considérese que Y_{ij} es una observación de una población $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ con $i = 1, \ldots, n_i$, y $j = 1, \ldots, q$. Los datos se pueden visualizar de la siguiente forma

Población	Muestra	Media muestral
1	$\boldsymbol{Y}_{11}, \boldsymbol{Y}_{12}, \ldots, \boldsymbol{Y}_{1n_1}$	$\overline{\boldsymbol{Y}}_{1\bullet}$
2	$\boldsymbol{Y}_{21}, \boldsymbol{Y}_{22}, \ldots, \boldsymbol{Y}_{2n_2}$	$\overline{\boldsymbol{Y}}_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$
q	$\boldsymbol{Y}_{q1}, \boldsymbol{Y}_{q2}, \ldots, \boldsymbol{Y}_{qn_q}$	$\overline{\boldsymbol{Y}}_{q\bullet}$

La media $\overline{\boldsymbol{Y}}_{i\bullet}$ en cada muestra se obtiene mediante

$$\overline{\boldsymbol{Y}}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{Y}_{ij} = \frac{1}{n_i} \boldsymbol{Y}_{i\bullet}, \text{ para } i = 1, \ldots, q.$$

La media general $\overline{\boldsymbol{Y}}_{\bullet\bullet}$ se obtiene de

$$\overline{\boldsymbol{Y}}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{q} \sum_{j=1}^{n_i} \boldsymbol{Y}_{ij} = \frac{1}{N} \sum_{i=1}^{q} \overline{\boldsymbol{Y}}_{i\bullet}$$

con $N = \sum_{i=1}^{q} n_i$, el número total de observaciones.
El modelo que relaciona las observaciones con los parámetros $\boldsymbol{\mu}_i$ es de la forma

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu}_i + \mathbf{E}_{ij}, \text{ con } \mathbf{E}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \text{ para } i = 1, \ldots, q \text{ y } j = 1, \ldots, n_i.$$

$$\begin{pmatrix} \boldsymbol{Y}'_{11} \\ \boldsymbol{Y}'_{12} \\ \vdots \\ \boldsymbol{Y}'_{1n_1} \\ -- \\ \vdots \\ -- \\ \boldsymbol{Y}'_{q1} \\ \boldsymbol{Y}'_{q2} \\ \vdots \\ \boldsymbol{Y}'_{qn_q} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_q} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \vdots \\ \boldsymbol{\mu}'_q \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}'_{11} \\ \boldsymbol{\varepsilon}'_{12} \\ \vdots \\ \boldsymbol{\varepsilon}'_{1n_1} \\ -- \\ \vdots \\ -- \\ \boldsymbol{\varepsilon}'_{q1} \\ \boldsymbol{\varepsilon}'_{q2} \\ \vdots \\ \boldsymbol{\varepsilon}'_{qn_q} \end{pmatrix}$$

$$\boldsymbol{Y} = \bigoplus_{i=1}^{q} \mathbf{1}_{n_i} \boldsymbol{\mu}_i + \mathbf{E}.$$

La hipótesis a verificar es la igualdad de los vectores de medias de las q–poblaciones; es decir,

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q. \tag{3.41}$$

La hipótesis planteada en (3.41) se puede escribir en la forma

$$\begin{aligned} H_0 : \ & \boldsymbol{\mu}_1 - \boldsymbol{\mu}_q = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_q = \cdots = \boldsymbol{\mu}_{q-1} - \boldsymbol{\mu}_q = \mathbf{0} \\ : \ & \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \vdots \\ \boldsymbol{\mu}'_q \end{pmatrix} = \mathbf{0}. \end{aligned}$$

La ecuación (3.39) se utiliza para contrastar esta hipótesis. La región de rechazo a un nivel de significación α es

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|} = \frac{|N\widehat{\boldsymbol{\Sigma}}|}{|N\widehat{\boldsymbol{\Sigma}}_\omega|} < \Lambda_{(\alpha, p, \nu_H, \nu_E)} \tag{3.42}$$

donde $\nu_H = q - 1$ son los grados de libertad para la hipótesis, $\nu_E = N - q$ son los grados de libertad del error ($N = \sum_{i=1}^{q} n_i$).

Más explícitamente

$$\boldsymbol{H} = \sum_{i=1}^{q} n_i \left(\overline{\boldsymbol{Y}}_{i\bullet} - \overline{\boldsymbol{Y}}_{\bullet\bullet})\left(\overline{\boldsymbol{Y}}_{i\bullet} - \overline{\boldsymbol{Y}}_{\bullet\bullet}\right)^T$$

$$\boldsymbol{E} = \sum_{i=1}^{q} \sum_{j=1}^{n_i} \left(\boldsymbol{Y}_{ij} - \overline{\boldsymbol{Y}}_{i\bullet}\right)\left(\boldsymbol{Y}_{ij} - \overline{\boldsymbol{Y}}_{i\bullet}\right)^T$$

$$\boldsymbol{E} + \boldsymbol{H} = \sum_{i=1}^{q} \sum_{j=1}^{n_i} \left(\boldsymbol{Y}_{ij} - \overline{\boldsymbol{Y}}_{\bullet\bullet}\right)\left(\boldsymbol{Y}_{ij} - \overline{\boldsymbol{Y}}_{\bullet\bullet}\right)^T. \tag{3.45}$$

Esta escritura de $\boldsymbol{E}$ permite encontrar un estimador insesgado de $\boldsymbol{\Sigma}$. De esta manera:

$$\boldsymbol{E} = (n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_q - 1)\boldsymbol{S}_q = \sum_{i=1}^{q}(n_i - 1)\boldsymbol{S}_i, \tag{3.46}$$

donde $\boldsymbol{S}_i$ es la matriz de covarianzas de la i-ésima muestra. Así, la matriz de varianzas y covarianzas estimada, puesto que las poblaciones se han considerado con igual matriz de covarianzas, es:

$$\boldsymbol{S}_p = \frac{1}{\sum_{i=1}^{q}(n_i - 1)}\boldsymbol{E} = \frac{\sum_{i=1}^{q}(n_i - 1)\boldsymbol{S}_i}{\sum_{i=1}^{q}(n_i - 1)}.$$

Es inmediato que para $p = 1$ (caso univarido), la razón de máxima verosimilitud se reduce a la conocida estadística F; así, se rechaza H_0 si:

$$\frac{\sum_i n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2}{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2}\left(\frac{N-q}{q-1}\right) > F_{(\alpha,q-1,N-q)}$$

La distribución exacta de Λ ha sido obtenida para algunos casos especiales, la tabla 3.8 los resume.

Para muestras de tamaño grande se tiene la estadística de Bartlett

$$V = -\left(N - 1 - \frac{(p+q)}{2}\right)\ln\Lambda$$

$$= -\left(N - 1 - \frac{(p+q)}{2}\right)\ln\left(\frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}\right),$$

la cual tiene aproximadamente una distribución Ji–cuadrado con $p(q-1)$ grados de libertad. Se rechaza H_0 para valores de V mayores que $\chi^2_{(\alpha,p(q-1))}$.

Tabla 3.8: Relación entre las estadísticas Λ y F

No. Variables	No. Grupos	Transformación	Distribución F
$p = 1$	$q \geq 2$	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-q}{q-1}\right)$	$F_{(q-1,N-q)}$
$p = 2$	$q \geq 2$	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{N-q-1}{q-1}\right)$	$F_{(2(q-1),2(N-q-1))}$
$p \geq 1$	$q = 2$	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-p-1}{p}\right)$	$F_{(p,N-p-1)}$
$p \geq 1$	$q = 3$	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{N-p-2}{p}\right)$	$F_{(2p,2(N-p-2))}$

Ejemplo 3.6.1. Con los siguientes datos se quiere establecer si tres métodos de enseñanza producen el mismo rendimiento promedio en matemáticas y escritura en niños de características similares.
Es éste un problema de análisis de varianza multivariado, con $p = 2$ que corresponde a los puntajes en matemáticas y escritura por estudiante. El número

Tabla 3.9: Datos de rendimiento bajo tres métodos de enseñanza

Método 1	$\binom{69}{75}$ $\binom{69}{70}$ $\binom{71}{73}$ $\binom{78}{82}$ $\binom{79}{81}$ $\binom{73}{75}$
Método 2	$\binom{69}{70}$ $\binom{68}{74}$ $\binom{75}{80}$ $\binom{78}{85}$ $\binom{68}{68}$ $\binom{63}{68}$ $\binom{72}{74}$ $\binom{63}{66}$ $\binom{71}{76}$ $\binom{72}{78}$ $\binom{71}{73}$ $\binom{70}{73}$ $\binom{56}{59}$ $\binom{77}{83}$
Método 3	$\binom{72}{79}$ $\binom{64}{65}$ $\binom{74}{74}$ $\binom{74}{74}$ $\binom{72}{75}$ $\binom{82}{84}$ $\binom{69}{68}$ $\binom{76}{76}$ $\binom{68}{65}$ $\binom{78}{79}$ $\binom{70}{71}$ $\binom{60}{61}$

Fuente: Freund, Litell & Spector (1986)

de poblaciones es $q = 3$; es decir, las tres metodologías. Los resultados del experimento se muestran en la tabla 3.9.
Se hará el análisis de varianza univariado (ANDEVA); es decir, para cada una de las dos variables, y el análisis de varianza multivariado (ANAVAMU) que se sugiere en este capítulo.
Las tablas 3.10 y 3.11 corresponden al análisis de varianza para cada una de las variables en forma separada.

Tabla 3.10: ANDEVA para matemáticas

F. de Variación	G. L.	S. C.	C. M.	Valor F	$Pr > F$
Métodos	2	60.6051	30.3025	0.91	0.4143
Error	28	932.8788	33.3171		
Total	30	993.4839			

De los resultados mostrados en la tabla 3.10 se puede afirmar que las metodologías no producen rendimientos promedios diferentes en matemáticas, en esta clase de niños.
Una conclusión similar se puede extraer de la tabla 3.11 para la variable escritura.

Tabla 3.11: ANDEVA para escritura

F. de Variación	G. L.	S. C.	C. M.	Valor F	$Pr > F$
Métodos	2	49.7359	24.8679	0.56	0.5776
Error	28	1243.9416	44.4265		
Total	30	1293.6775			

$$\boldsymbol{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij} \quad \text{con } i = 1, 2 \text{ y } 3 \quad j = 1, \ldots, n_i.$$

En este caso $n_1 = 6$, $n_2 = 14$, $n_3 = 11$ y $N = 31$.
Mediante la hipótesis nula se afirma que los métodos producen un rendimiento en promedio igual en matemáticas y en escritura; es decir,

$$H_0 : \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3.$$

Las matrices de sumas de cuadrados (covariabilidad) dentro y entre tratamientos se obtienen aplicando (3.45)

$$\boldsymbol{E} = \begin{pmatrix} 932.87879 & 1018.6818 \\ 1018.6818 & 1243.9416 \end{pmatrix} \qquad \boldsymbol{H} = \begin{pmatrix} 60.6050 & 31.5117 \\ 31.5117 & 49.7358 \end{pmatrix}.$$

El valor del lambda de Wilks es

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|} = \frac{\begin{vmatrix} 932.8788 & 1018.6818 \\ 1018.6818 & 1243.9416 \end{vmatrix}}{\begin{vmatrix} 993.4838 & 1050.1935 \\ 1050.1935 & 1293.6774 \end{vmatrix}} = 0.6731.$$

De la tabla 3.8 y como $p = 2$ y $q = 3$, se puede utilizar la estadística

$$\left(\frac{1 - \Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{n - q - 1}{q - 1}\right) = \left(\frac{1 - \sqrt{0.6731}}{\sqrt{0.6731}}\right)\left(\frac{31 - 3 - 1}{3 - 1}\right)$$
$$= 2.954851.$$

El valor anterior comparado con $F_{(5\%,2(3-1),2(31-3-1))} = F_{(5\%,4,54)} \approx 2.5$ (tabla C.8), permite afirmar que el puntaje promedio no es el mismo para las tres metodologías. ¡*El resultado no es el mismo que se obtuvo con los análisis de varianza univariados*!