ERRAMIENTAS DEL ÁLGEBRA MATRICIAL PARA ABORDAR LOS MÉTODOS ESTADÍSTICOS MULTIVARIANTES

Gudberto José León Rangel
Universidad de Los Andes, Venezuela
Departamento de Estadística de la Facultad de Ciencias Económicas y Sociales (FACES).
Director de la Escuela de Estadística y Coordinador de la Comisión Curricular de la carrera de Estadística.
Correo electrónico: gudberto@ula.ve

Introducción

El Análisis Multivariante está constituido por un conjunto de métodos y técnicas utilizadas en el estudio del comportamiento simultáneo de varias variables. Permite obtener una visión de conjunto de fenómenos de la realidad cuya complejidad exige que sean estudiados con técnicas de mayor alcance que las de la estadística univariante o bivariante. Su objetivo fundamental es resumir y sintetizar la información contenida en grandes conjuntos de datos, con el fin de lograr una mejor comprensión del fenómeno en estudio.

El término multivariante (del inglés *multivariate*) se refiere precisamente al hecho de que se consideran múltiples variables. Hasta no hace muchos años, los métodos multivariantes habían permanecido en el campo meramente teórico. Con el uso actual de los potentes equipos de computación, estos métodos son utilizados en la mayoría de las investigaciones científicas, habiéndose comprobado ampliamente su eficacia en el tratamiento de grandes masas de datos.

El término Análisis de Datos surge en la década de 1960 con la intención de distinguirlo del análisis multivariante clásico basado en modelos y supuestos teóricos, con la idea de enfatizar la descripción de conjuntos numerosos de datos.

En la actualidad se reconoce y aprecia la importancia de la estadística en todas las esferas de la ciencia. Los métodos estadísticos se usan incluso en disciplinas tales como historia, literatura y lingüística, en las cuales la idea de realizar estudios cuantitativos era inconcebible hasta hace unos pocos años. Actualmente son muchos los problemas cuya solución puede lograrse más fácilmente, o incluso únicamente, con la ayuda de los métodos multivariantes, en campos tan diversos como: agricultura, antropología, física, educación, economía, análisis de mercados, medicina, psicología, sociología y biología.

El origen del análisis multivariante descansa sobre los conceptos matemáticos desarrollados por matemáticos franceses e italianos del siglo pasado, quienes se dedicaron a estudiar los aspectos del álgebra matricial que sirvieron de base para la factorización de una matriz en sus valores y vectores singulares. Los primeros estudios multivariantes se remontan a las generalizaciones de las investigaciones sobre correlación y regresión realizadas a principios del siglo XX por Francis Galton, Karl Pearson y Charles Spearman, científicos ingleses que trabajaban en Sicología y Biometría.

Caracterización de muestras de poblaciones multivariantes

Nociones fundamentales del álgebra matricial permiten abordar de manera simplificada el cálculo de las medidas resumen, frecuentemente utilizadas, que ilustran los principales aspectos de la información contenida en los arreglos de datos.

Organización de datos multivariantes

Generalmente, la información sobre la que se aplican los métodos multivariantes se organiza sobre una matriz de datos $\boldsymbol{X}$ con n filas y p columnas.

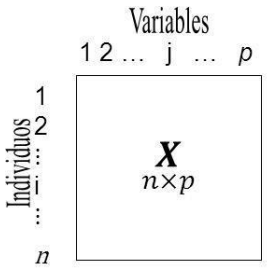
$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \square & \vdots & \square & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \square & \vdots & \square & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

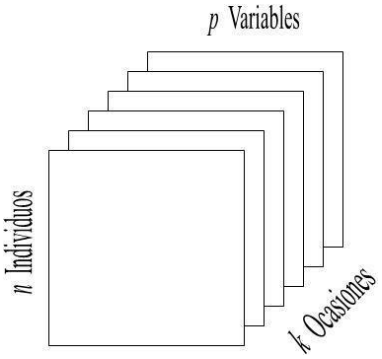
Variable j ↓

→ Individuo i

Datos Multivía

En este caso, el arreglo de la información se presenta en tres o más vías:



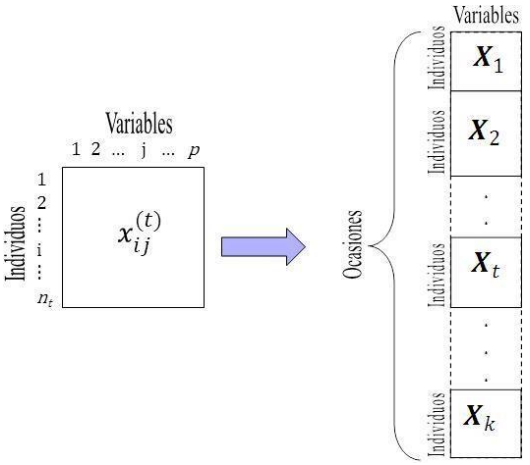


Datos de tres vías:

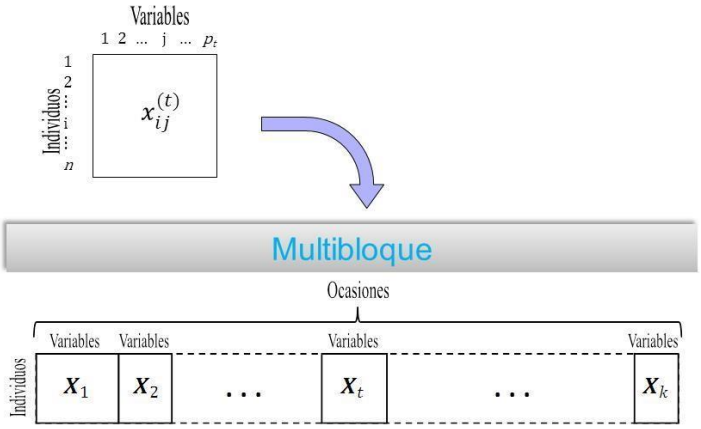
Describen los mismos individuos y las mismas variables en cada ocasión.

Datos de Conjuntos Múltiples:

Se pueden presentar de dos maneras, las mismas variables para diferentes conjuntos de individuos:

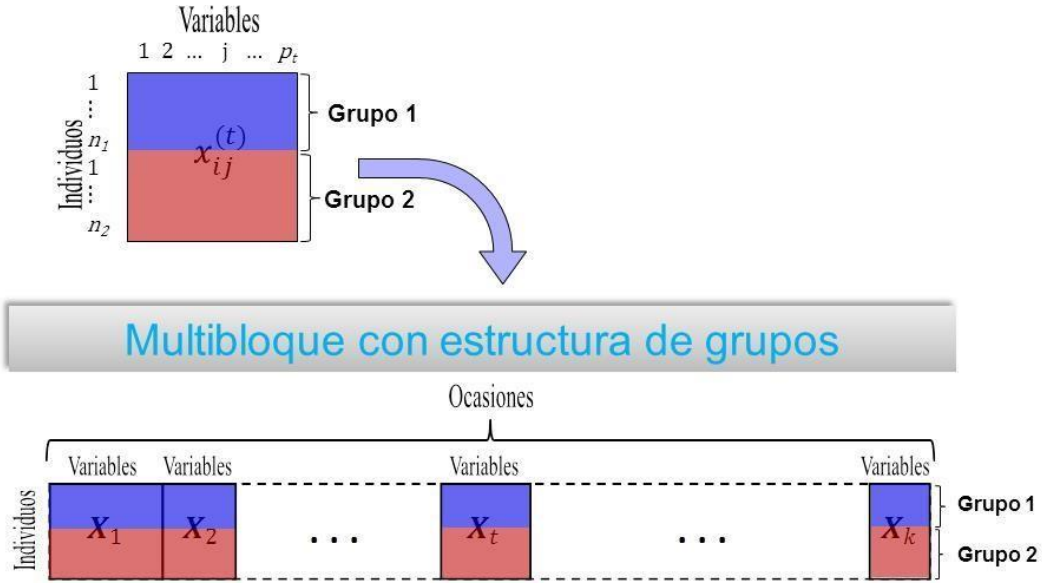


O como los mismos individuos y diferentes conjuntos de variables:



Datos Multivía con Estructura de Grupos

En algunas investigaciones interesa la situación en que los mismos individuos han sido caracterizados por *k* conjuntos de variables (un multibloque) y esos individuos se presentan clasificados en dos grupos.



Matriz de datos

Como se mencionó antes, los datos recogidos en una investigación pueden arreglarse en una matriz:

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

El i-ésimo vector fila de la matriz de datos contiene las observaciones correspondientes al individuo i en cada una de las p variables:

$$X_i^t = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{ij} \quad \cdots \quad x_{ip})$$

El j-ésimo vector columna describe la información de la variable j medida sobre los n individuos:

$$X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_n \end{pmatrix}$$

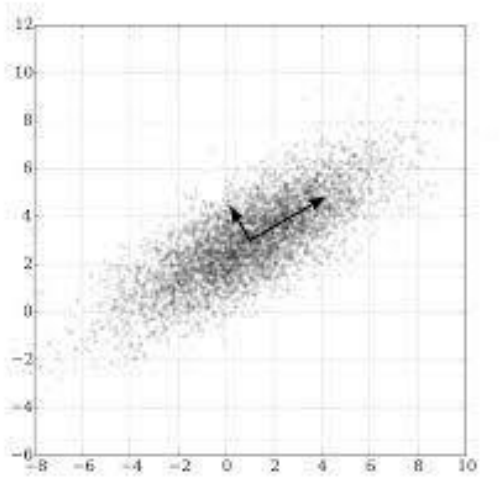
Ejemplo:

Al organizar sobre un arreglo matricial la información correspondiente a las edades y pesos de n=10 niños, se obtiene la siguiente matriz de datos:

		Edad		Peso
Niños	1	7	⋮	25
	2	10	⋮	28
	3	12	⋮	36
	4	4	⋮	17
	5	5	⋮	21
	6	13	⋮	48
	7	7	⋮	28
	8	10	⋮	32
	9	8	⋮	27
	10	11	⋮	38

Nube de datos

El gráfico de dispersión para dos variables en una matriz de datos puede presentar una forma como la siguiente:



Ejemplo: (Ramírez y Vásquez, 2006)

Los resultados obtenidos por doce (12) maestros que aspiran ocupar puestos de trabajo del Ministerio de Educación, en seis pruebas de aptitud medidas en una escala del 1 al 20 son:

	X1	X2	X3	X4	X5	X6
Maestro1	16	13	12	17	12	17
Maestro2	18	17	11	17	15	17
Maestro3	15	15	14	14	16	16
Maestro4	14	16	15	13	17	14
Maestro5	16	19	13	14	18	15
Maestro6	17	15	12	17	14	19
Maestro7	11	16	17	11	18	12
Maestro8	16	17	13	16	16	15
Maestro9	16	19	14	13	18	12
Maestro10	16	16	14	15	16	15
Maestro11	11	19	18	9	20	10
Maestro12	15	14	14	15	15	18

Donde las variables representan los resultados de las pruebas de aptitud en las siguientes áreas:

X_1: Inglés
X_2: Habilidad numérica
X_3: Ciencias naturales
X_4: Cultura general
X_5: Capacidad de abstracción
X_6: Castellano

Matrices Grammian

Se definen como matrices Grammian asociadas a la matriz $\boldsymbol{X}$ a las matrices: $\boldsymbol{X^tX}$ y $\boldsymbol{XX^t}$

Estas matrices son de interés en estadística, sobre ellas se organiza información fundamental para:

- El análisis de las relaciones entre variables
- El análisis del parecido entre individuos

La matriz de varianzas y covarianzas, así como la matriz de correlaciones son ejemplos de este tipo de matrices.

Información estadística en una matriz de datos

Al efectuar operaciones matriciales sobre una matriz de datos $\boldsymbol{X}_{n\times p}$, que contiene la información obtenida al caracterizar n individuos de acuerdo con p variables, es posible obtener información estadística como:

1. **Vector de medias o centro de gravedad de las p variables**:

Sea el vector fila $j^t = (1,1,\dots,1)$, entonces,

$$\bar{X}^t = \frac{1}{n} j^t \boldsymbol{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$$

donde:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

es la media aritmética de la variable j.

Efectivamente,

$$\bar{X}^t = \frac{1}{n} j^t \boldsymbol{X}$$

$$= \left(\frac{1}{n} j^t X^1, \frac{1}{n} j^t X^2, \dots, \frac{1}{n} j^t X^p\right)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_{i1}, \frac{1}{n}\sum_{i=1}^{n} x_{i2}, \dots, \frac{1}{n}\sum_{i=1}^{n} x_{ip}\right)$$

$$= (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$$

Ejemplo:

Calculando el vector de medias sobre la matriz de datos del ejemplo anterior, se obtiene:

	X1	X2	X3	X4	X5	X6
Promedio	**15,1**	**16,3**	**13,9**	**14,3**	**16,3**	**15,0**
Desv. Estand.	2,2	2,0	2,0	2,5	2,1	2,7
Coef. Var.	0,14	0,12	0,15	0,17	0,13	0,18

En la tabla anterior también se muestra la desviación estándar y el coeficiente de variación para cada una de las variables.

Como todas las variables están medidas en las mismas unidades y sus niveles de dispersión son similares, es posible considerar que los promedios producen una ordenación de las variables indicando en donde se obtuvieron los puntajes más y menos elevados.

1. **Matriz centrada por columna (o de desvíos)**

Los *datos centrados* $\boldsymbol{X}_c = (x_{ij} - \bar{x}_j)$:

- Dan un valor agregado a la información reportada por los datos originales.
- Capta la magnitud en la que el valor observado del individuo i en la variable j se aleja de la correspondiente media.
- El signo resultante indica la dirección del alejamiento La matriz centrada se puede escribir en la forma:

$$\boldsymbol{X}_c = \left(\boldsymbol{I} - \frac{1}{n} jj^t\right)\boldsymbol{X}$$

$$= (x_{ij} - \bar{x}_j)$$

donde $\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)$ se conoce en la literatura como *matriz de centraje*.

- $\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)$ es simétrica: $\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)^{\boldsymbol{t}}=\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)$
- $\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)$ es idempotente: $\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)=\left(\boldsymbol{I}-\frac{1}{n}jj^t\right)$

La matriz centrada constituye la base para la obtención de la matriz de varianzas y covarianzas.

Matriz de datos centrada:

Los datos centrados se pueden organizar en una matriz de datos:

$$\boldsymbol{X}_c=\begin{pmatrix} x_{11}-\overline{x}_1 & \cdots & x_{1j}-\overline{x}_j & \cdots & x_{1p}-\overline{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{i1}-\overline{x}_1 & \cdots & x_{ij}-\overline{x}_j & \cdots & x_{ip}-\overline{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{n1}-\overline{x}_1 & \cdots & x_{nj}-\overline{x}_j & \cdots & x_{np}-\overline{x}_p \end{pmatrix}$$

Ejemplo:

La matriz centrada de los datos referentes a las seis pruebas de aptitud mencionadas en los ejemplos anteriores está dada por:

X1	X2	X3	X4	X5	X6
0,9167	-3,3333	-1,9167	2,7500	-4,2500	2,0000
2,9167	0,6667	-2,9167	2,7500	-1,2500	2,0000
-0,0833	-1,3333	0,0833	-0,2500	-0,2500	1,0000
-1,0833	-0,3333	1,0833	-1,2500	0,7500	-1,0000
0,9167	2,6667	-0,9167	-0,2500	1,7500	0,0000
1,9167	-1,3333	-1,9167	2,7500	-2,2500	4,0000
-4,0833	-0,3333	3,0833	-3,2500	1,7500	-3,0000
0,9167	0,6667	-0,9167	1,7500	-0,2500	0,0000
0,9167	2,6667	0,0833	-1,2500	1,7500	-3,0000
0,9167	-0,3333	0,0833	0,7500	-0,2500	0,0000
-4,0833	2,6667	4,0833	-5,2500	3,7500	-5,0000
-0,0833	-2,3333	0,0833	0,7500	-1,2500	3,0000

2. Matriz de varianzas y covarianzas

En los métodos estadísticos multivariante, el interés se centra más en el análisis de las varianzas y covarianzas, que en el estudio de los valores de las variables en sí mismos.

La matriz de varianzas y covarianzas puede escribirse:

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}_c^t\boldsymbol{X}_c = (s_{jk})$$

$$\boldsymbol{S}_{p\times p} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2k} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jk} & \cdots & s_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{p1} & s_{n2} & \cdots & s_{pk} & \cdots & s_{pp} \end{pmatrix}$$

El término genérico de la matriz S:

$$s_{jk} = \frac{1}{n-1}\left(x_{1j}-\overline{x}_j, x_{2j}-\overline{x}_j, \ldots, x_{nj}-\overline{x}_j\right)\begin{pmatrix} x_{1k}-\overline{x}_k \\ x_{2k}-\overline{x}_k \\ \vdots \\ x_{nk}-\overline{x}_k \end{pmatrix}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)\left(x_{ik}-\overline{x}_k\right)$$

,

El cual representa la covarianza entre las variables j y k.

Los elementos sobre la diagonal principal de la matriz $\boldsymbol{S}$, son las varianzas de las p variables de la matriz de datos $\boldsymbol{X}$.

$$s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)\left(x_{ij}-\overline{x}_j\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)^2$$

La matriz de varianzas y covarianzas de los datos del ejemplo que estamos siguiendo, está dada por:

4,6288	-0,6667	-4,1742	4,7955	-2,9318	4,1818
-0,6667	3,8788	1,2121	-2,5455	3,4545	-3,5455
-4,1742	1,2121	4,0833	-4,7955	3,2955	-4,3636
4,7955	-2,5455	-4,7955	6,2045	-4,7045	5,9091
-2,9318	3,4545	3,2955	-4,7045	4,5682	-4,9091
4,1818	-3,5455	-4,3636	5,9091	-4,9091	7,0909

3. **Matriz de datos estandarizada**

$$\boldsymbol{X_e = X_c D_s^{-1}} = \left(\frac{x_{ij} - \overline{x}_j}{s_j}\right)$$

donde $\boldsymbol{D_s}$ es una matriz diagonal que contiene las desviaciones estándar de las p variables:

$$\boldsymbol{D_s} = diag(s_1, s_2, \ldots, s_p)$$

Así,

$$\boldsymbol{D_s^{-1}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{s_j} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \frac{1}{s_p} \end{pmatrix}$$

$\frac{x_{ij}-\overline{x}_j}{s_j}$ El término mide el alejamiento de la observación original respecto de la media, expresado en unidades de desviación estándar. El valor original obtenido por un individuo en una variable no proporciona información sobre su posición con respecto al resto de las observaciones. Mientras que la estandarización ubica al dato en referencia al conjunto total al cual pertenece. Así se pueden hacer comparaciones objetivas sobre los valores observados de la variable.

Ejemplo:

La matriz de datos estandarizada para nuestros datos es:

X1	X2	X3	X4	X5	X6
0,4261	-1,6925	-0,9485	1,1040	-1,9885	0,7511
1,3557	0,3385	-1,4434	1,1040	-0,5848	0,7511
-0,0387	-0,6770	0,0412	-0,1004	-0,1170	0,3755
-0,5035	-0,1693	0,5361	-0,5018	0,3509	-0,3755
0,4261	1,3540	-0,4536	-0,1004	0,8188	0,0000
0,8909	-0,6770	-0,9485	1,1040	-1,0527	1,5021
-1,8979	-0,1693	1,5259	-1,3048	0,8188	-1,1266
0,4261	0,3385	-0,4536	0,7026	-0,1170	0,0000
0,4261	1,3540	0,0412	-0,5018	0,8188	-1,1266
0,4261	-0,1693	0,0412	0,3011	-0,1170	0,0000
-1,8979	1,3540	2,0207	-2,1077	1,7545	-1,8777

-0,0387	-1,1848	0,0412	0,3011	-0,5848	1,1266

4. Matriz de correlaciones

$$\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{X}_e^t\boldsymbol{X}_e = (r_{jk})$$

Nótese que la matriz de correlaciones es una matriz de varianzas y covarianzas calculada sobre datos estandarizados.

También puede expresarse en la forma:

$$\boldsymbol{R} = \boldsymbol{D}_s^{-1}\boldsymbol{S}\boldsymbol{D}_s^{-1}$$

El término genérico es:

$$r_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{ij}-\overline{x}_j}{s_j}\right)\left(\frac{x_{ik}-\overline{x}_k}{s_k}\right)$$

$$= \frac{s_{jk}}{s_j\, s_k}$$

- La información sobre los coeficientes de correlación de Pearson entre dos variables se registra en la matriz de correlaciones.
- El grado y dirección de dependencia lineal entre variables, dos a dos, es medido por r_{jk}.

La traza como medida de variabilidad total en datos multivariantes

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2 \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n} d^2(x_{ij}, \overline{x}_j)$$

Siendo $d^2(x_{ij}, \overline{x}_j)$ el cuadrado de la distancia euclídea unidimensional entre el valor x_{ij} y la media $\overline{x}_j$.

La traza de la matriz de varianzas y covarianzas, es una generalización del concepto de varianza de una variable sobre el espacio $p$variante $\mathbb{R}^p$.

Consideremos los valores observados de la variable X^j sobre $n$individuos:

$$X^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_n \end{pmatrix}$$

Calculando la varianza:

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2$$

La varianza constituye una medida global promedio del parecido entre cada una de las observaciones y el valor que mejor los representa ($\overline{x}_j$) en el sentido de los mínimos cuadrados:

Consideremos ahora:

$$\frac{1}{\sqrt{n-1}}\boldsymbol{X}_c = \frac{1}{\sqrt{n-1}}\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1j}-\bar{x}_j & \cdots & x_{1p}-\bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{i1}-\bar{x}_1 & \cdots & x_{ij}-\bar{x}_j & \cdots & x_{ip}-\bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & \cdots & x_{nj}-\bar{x}_j & \cdots & x_{np}-\bar{x}_p \end{pmatrix}$$

Si se suman los cuadrados de los elementos de la i-ésima fila se obtiene, salvo el factor $1/(n-1)$, la distancia entre el individuo X_i y el vector de medias de las $p$variables.

La varianza también constituye una medida del parecido del perfil de valores de ese individuo respecto del perfil promedio

$$\frac{1}{n-1}d^2(X_i-\bar{X}) = \frac{1}{n-1}\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)^2$$

$$\frac{1}{\sqrt{n-1}}\boldsymbol{X}_c = \frac{1}{\sqrt{n-1}}\begin{pmatrix} x_{11}-\bar{x}_1 & \cdots & x_{1j}-\bar{x}_j & \cdots & x_{1p}-\bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{i1}-\bar{x}_1 & \cdots & x_{ij}-\bar{x}_j & \cdots & x_{ip}-\bar{x}_p \\ \vdots & & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & \cdots & x_{nj}-\bar{x}_j & \cdots & x_{np}-\bar{x}_p \end{pmatrix}$$

Al sumar los cuadrados de los elementos en la j-ésima columna, se obtiene la varianza de la variable j:

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

De los resultados anteriores se establece la relación fundamental:

$$\frac{1}{n-1}\sum_{i=1}^{n}d^2(X_i-\bar{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)$$

$$= \sum_{j=1}^{p}s_j^2 = tr(\boldsymbol{S})$$

$$\boldsymbol{S}_{p\times p} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_j^2 & \cdots & s_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{p1} & s_{n2} & \cdots & s_{pj} & \cdots & s_p^2 \end{pmatrix}$$

La traza de la matriz de varianzas y covarianzas constituye una medida fundamental de la variabilidad total multivariante.

De esta manera hemos visto como con operaciones básicas del álgebra matricial se han obtenidos medidas multivariantes que son muy útiles en estadística para la caracterización de una muestra.

Referencias

Johnson, A. y Wichern, D. (2007). *Applied multivaruate statistical analysis* (6a.ed.). Saddle River, NJ: Pearson Prentice Hall.

Lebart L., Morineau A., Warwick K. (1984). *Multivariate descriptive statistical analysis* (E. Moraillon, Trad.). Nueva York: Wiley. (Trabajo original publicado en 1977).

Ramírez, G. y Vásquez, M. (2008). *Análisis de Datos*. Universidad Central de Venezuela, Caracas.

Ramírez, G. y Vásquez, M. (2006). *Aspectos teóricos del álgebra matricial con aplicaciones estadísticas*. Universidad Central de Venezuela, Caracas.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2a. ed.). New York: Wiley.

Uriel, E. y Aldás, J. (2005). *Análisis multivariante aplicado*. Madrid: Thomson.