Estadística multivariada

Jaime Emmanuel Alcalá Temores

CEIC, UdeG

30 de enero de 2018

Introducción

En la estadística multivariable es un conjunto de métodos para describir (e interpretar) datos que provienen de la observación de más de una variable, de forma simultánea. Las variables se representan como una matriz $\mathbf{X} = (X_1, ...X_p)$, donde cada uno de sus componentes es un vector columna de la forma $X_i = (x_{i1}, x_{i2}..., x_{in})^T$.
En forma matricial, representamos los n individuos por p variables como la matriz $n \times p, \mathbf{X}$ como

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Ejemplo

Cuadro 1: Pesos cuerpo-cerebro

	X_1	X_2
Zorro blanco	3.38	44.5
Búho	0.48	15.5
Castor	1.35	8.1
Vaca	465	423
Lobo gris	36.33	119.5
Cabra	27.66	115
Corzo	14.83	98.2
Cobaya	1.04	5.5
Vervet	4.19	58
Chinchilla	0.43	6.4
Ardilla	0.1	4
Ardilla ártica	0.92	5.7
Rata africana	1	6.6
Musaraña	0	0.14

X_1: Peso corporal (kg)
X_2: Peso del cerebro (g)

Medidas de tendencia central

Media

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Varianza

$$S^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Desviación estándar

$$S = \sqrt{S^2}$$

Covarianza

$$S_{jk} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)$$

Coeficiente de correlación

$$r_{jk} = \frac{S_{jk}}{S_j S_k}$$

Estandarización de matrices

Si nuestras variables tienen diferentes unidades o si queremos que cada una de ellas tenga el mismo peso, debemos estandarizarlas. Estandarizar las variables significa extraer a cada variable la media y dividir eso por la desviación estándar.

$$Z_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i}$$

Que nos da la matriz

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1p} \\ z_{21} & z_{22} & \dots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{np} \end{bmatrix}$$

Ejemplo

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ 10 & 15 & 21 & 12 \\ 15 & 16 & 29 & 16 \\ 20 & 18 & 16 & 25 \\ 30 & 25 & 19 & 30 \end{bmatrix}$$

Promedios

$$\bar{X}_1 = 18.75, \bar{X}_2 = 18.5, \bar{X}_3 = 21.25, \bar{X}_4 = 20.75$$

Desviaciones estándar

$$\bar{S}_{X_1} = 8.54, \bar{S}_{X_2} = 4.51, \bar{S}_{X_3} = 5.56, \bar{S}_{X_4} = 8.22$$

Ejemplo

Matriz estandarizada

$$\mathbf{Z} = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 \\ -1.024695 & -0.7761823 & -0.6012934 & -1.0643593 \\ -0.439155 & -0.5544160 & 1.1479238 & -0.5777950 \\ 0.146385 & -0.1108832 & 0.4919673 & 0.5169745 \\ 1.317465 & 1.4414815 & -1.0385977 & 1.1251798 \end{bmatrix}$$

En la matriz estandarizada el promedio de cada columna es 0 y su desviación estándar 1.

Matriz de Varianza-Covarianza

Se denota por Σ.

$$\Sigma = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \dots & S_{np} \end{bmatrix}$$

Haciendo $X_0 = (X_{ij} - X_j)$, Σ se puede calcular mediante:

$$\Sigma = \left(\frac{1}{n-1}\right)\mathbf{X}_0'\mathbf{X}_0$$

En donde $\mathbf{X}_0'$ es la matriz transpuesta de $\mathbf{X}_0$. Σ es una matriz simétrica. La diagonal, $diag(\Sigma)$ son las varianzas.

Ejemplo

Con la matriz

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ 10 & 15 & 21 & 12 \\ 15 & 16 & 29 & 16 \\ 20 & 18 & 16 & 25 \\ 30 & 25 & 19 & 30 \end{bmatrix}$$

Centrando en la media para obtener $\mathbf{X}_0$, tenemos la matriz de varianza convarianza:

$$\Sigma = \begin{bmatrix} 72.9 & 37.5 & -15.42 & 67.92 \\ 37.5 & 20.3 & -11.8 & 33.5 \\ -15.42 & -11.8 & 20.92 & -11.75 \\ 67.92 & 33.5 & -11.75 & 67.59 \end{bmatrix}$$

Matriz de correlaciones

R_{jk} es la correlación de la variable X_j con la variable X_k

$$\mathbf{R} = \begin{bmatrix} 1 & R_{12} & \dots & R_{1p} \\ R_{21} & 1 & \dots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{n1} & R_{n2} & \dots & 1 \end{bmatrix}$$

En la diagonal tenemos unos, que es la correlación de X_j consigo misma. $\mathbf{R}$ se puede computar mediante

$$\mathbf{R} = \frac{\left(\frac{1}{n-1}\right) \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)}{S_j S_k}$$

O alternativamente como la matriz var-cov de Z:

$$\mathbf{R} = \left(\frac{1}{n-1}\right)\mathbf{Z}'\mathbf{Z}$$

Con

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ 10 & 15 & 21 & 12 \\ 15 & 16 & 29 & 16 \\ 20 & 18 & 16 & 25 \\ 30 & 25 & 19 & 30 \end{bmatrix}$$

Tenemos una matriz $\mathbf{R}$

$$\mathbf{R} = \begin{bmatrix} 1 & 0.97 & -0.34 & 0.97 \\ 0.97 & 1 & -0.57 & 0.9 \\ -0.39 & -0.57 & 1 & -0.3 \\ 0.97 & 0.9 & -0.31 & 1 \end{bmatrix}$$

Eigenvalores & Eigenvectores

Para una matriz cuadrada $n \times n$, $\mathbf{A}$, si existe un λ que satisfaga

$$\mathbf{Ax} = \lambda\mathbf{x}$$

Se dice que λ es un eigenvalor o valor propio correspondiente al vector propio $\mathbf{x}$. La suma de los eigenvalores de Σ, que corresponde a la suma de las varianzas, nos retorna la variación total

$$\sum_{j=1}^{p} s_j^2 = s_1^2 + s_2^2 + \cdots + s_p^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{j=1}^{p} \lambda_j$$