Tratamiento matricial de los datos multivariantes

Eliseo Martínez H.

1. Introducción

Intentaremos conciliar el lenguaje matricial con el lenguaje coloquial de cuestionario en que se hacen p preguntas a n personas, y suponiendo que cada respuesta es numérica o es fácilmente cuantificable, por ejemplo cuando se pregunta por el género. Vamos a suponer entonces que tenemos n individuos (o en términos estadísticos n unidades muestrales), donde a cada individuo le haremos p preguntas (de otra forma, se le medirán p atributos cuantificables o atributos cuantitativos ordinales). Supongamos que las respuestas numéricas las ubicamos en el siguiente arreglo bidimensional,

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

en forma más compacta, $X = (x_{ij})_{n\times p}$, $i = 1, ..., n;\ j = 1, ..., p$. Entendiendo que x_{ij} es la respuesta a la pregunta j-ésima realizada por el individuo i-ésimo. La fila de respuestas realizada por el individuo i, matricialmente la denotamos como[1]

$$\mathbf{x}_i^t = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$$

De tal manera que

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix}$$

Es de gran importancia obtener la el valor medio o promedio de las respuestas a la pregunta j realizada por los n individuos, esto es necesitamoe el cálculo de

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij};\ \ j = 1, \cdots, p$$

[1] Los vectores filas los consideraremos en el orden traspuesto, toda vez que los vectores, para nosotros, serán siempre vectores columnas.

las p medias. Con estos p valores formamos el vector de medias, esto es

$$\overline{\mathbf{x}} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix}$$

Puesto que tenemos que

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}; \ \ i = 1, \cdots, n$$

la i-ésima fila de la matriz $\mathbf{X}$, puesta ahora como vector columna, no resulta complicado verificar que el vector de medias se puede obtener también de la siguiente manera

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

No obstante la mejor representación para el vector de medias es a través de la propia matriz de datos $\mathbf{X}$. En efecto, se verifica que

$$\overline{\mathbf{x}} = \frac{1}{n} \mathbf{X}^t \mathbf{1} \tag{1}$$

donde

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

es el vector de dimensión n y cuyas entradas son "unos".

Observe la matriz de datos $\mathbf{X}$, si sumamos a través de las n filas, obtendremos para cada columna j un total, y este total será igual, como es obvio, a n veces su respectivo promedio, de otra forma

$$\sum_{i=}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) = 0$$

2. Matriz de varianzas y covarianzas

Si observamos nuevamente la matriz de datos,

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

y considerando que cada columna es una variable observada, en rigor n respuestas a la

misma pregunta, podemos estudiar la covarianza entre diferentes variables (respuestas a diferentes preguntas). Definamos entonces la varianza entre la columna j y la columna k como[2]

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

Esta covarianza medirá la dependencia lineal entre ambas variables. Observemos que si $j = k$, entonces obtenemos la varianza de la j-ésima variable (la columna j), esto es s_j^2. Todas las varianzas y covarianzas nos permiten definir la llamada **matriz de varianza y covarianza** como sigue

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

que es una matriz cuadrada de orden $p \times p$ simétrica.

Esta matriz $\mathbf{S}$ la podemos calcular directamente de la matriz de datos $\mathbf{X}$. En efecto, definamos la *matriz de datos centrada*, como la matrriz de datos al cual a cada columna le restamos la media respectiva de dicha columna, esto es

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}^t$$

en forma más detallada

$$\widetilde{\mathbf{X}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} - \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \vdots & \vdots & \cdots & \vdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \end{pmatrix}$$

Sustituyendo el vector de medias por su expresión dada en (2), obtenemos

$$\begin{aligned} \widetilde{\mathbf{X}} &= \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^t\mathbf{X} \\ &= (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^t) \cdot \mathbf{X} \\ &= \mathbf{P} \cdot \mathbf{X} \end{aligned}$$

donde la matriz $\mathbf{P}$ está definida por

$$\mathbf{P} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^t)$$

y resulta ser una matriz simétrica de orden $n \times n$, idempotente y de rango $n - 1$. Viéndola con más detalle

$$\mathbf{P} = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1 - \frac{1}{n} \end{pmatrix}$$

[2] En algunos libros se considera $s_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$

Entonces la matriz de varianza y covarianza puede escribirse como

$$\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$$

3. Propiedades de la matriz de varianza y covarianza

La varianza de una variable unidimensional siempre es positiva, en nuestro caso para datos multivariantes la situación es similar. Esto es la matriz de varianza y covarianza $\mathbf{S}$ es semi-definida positiva. Es decir que para todo vector en $\mathbf{y} \in \mathbb{R}^p$ se satisface que $\mathbf{y}^t\mathbf{S}\,\mathbf{y} \geq 0$. Vamos a demostrar esto.

Sea $\mathbf{w}$ cualquier vector de dimensión p, definamos la variable unidimensional

$$v_i = \mathbf{w}^t(\mathbf{x}_i - \overline{\mathbf{x}}) \tag{2}$$

Observemos que el vector $\mathbf{x}_i - \overline{\mathbf{x}}$ es la i-ésima fila de la matriz de datos en que a cada componente se le resta la media de cada columna de la matriz de datos. La media de los valores v_i es:

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{n}\ \mathbf{w}^t \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}}) = 0$$

y entonces su varianza, que es no negativa, es:

$$\begin{aligned} Var\ (v) &= \frac{1}{n}\sum_{i=1}^{n} v_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{w}^t(\mathbf{x}_i - \overline{\mathbf{x}})\right]\left[(\mathbf{x}_i - \overline{\mathbf{x}})^t\mathbf{w}\right] \geq 0 \\ &= \mathbf{w}^t S\,\mathbf{w} \geq 0 \end{aligned}$$

Y puesto que $\mathbf{w}$ es cualquier vector, entonces se concluye que $\mathbf{S}$ es semidefinida positiva. Además supongamos que λ_i es un autovalor de S, es decir que existe un $\mathbf{w}_i$ tal que $\mathbf{S}\,\mathbf{w}_i = \lambda_i\mathbf{w}_i$, entonces $\mathbf{w}_i^t\,\mathbf{S}\,\mathbf{w}_i = \mathbf{w}_i^t\lambda_i\mathbf{w}_i \geq 0$, y esto significa que $\lambda_i \geq 0$. Es decir, todos los autovalores de $\mathbf{S}$ son no negativos.

Supongamos ahora que la matriz $\mathbf{S}$ es singular. Si este es el caso entonces existe un vector $\mathbf{w}$ tal que satisface la igualdad $\mathbf{w}^t\,\mathbf{S}\,\mathbf{w} = 0$. De tal manera que si definimos las variables v_i como en (2) estas variables tendrán varianza nula, y puesto que su media es ceo, entonces estas variables toman el valor cero. En consecuencia, para cualquier i (cualquier fila) se satisface que

$$\sum_{j=1}^{p} w_j(x_{ij} - \overline{x}_j) = 0 \quad \forall\, i$$

Y esta ecuación implica que las p variables de la fila i no son independientes, puesto que podemos despejar cualquier x_{ij} en función de los restantes, por ejemplo x_{i1}:

$$x_{i1} = \overline{x}_1 - \frac{w_2}{w_1}(x_{i2} - \overline{x}_2) - \cdots - \frac{w_p}{w_1}(x_{ip} - \overline{x}_p)$$

entendiendo que $w_1 \neq 0$.

Por lo tanto, si existe algún vector $\mathbf{w}$ para el cual $\mathbf{w}^t\,\mathbf{S}\,\mathbf{w} = 0$, existe una relación lineal entre las variables (en nuestro lenguaje, una columna de respuestas es linealmente dependiente de las restantes columnas de respuestas, de otra forma una pregunta o atributo

tiene una relación lineal con las restantes preguntas o atributos).

El recíproco también es cierto, esto es si hay una relación lineal entre las variables, entonces podemos escribir $\mathbf{w}^t(\mathbf{x}_i - \overline{\mathbf{x}}) = 0$ para todo i, para algún $\mathbf{w}$ con componentes no todas nulas, es decir

$$\widetilde{\mathbf{X}}\,\mathbf{w} = \mathbf{0}$$

multiplicando esta expresión por la derecha por la matriz $\widetilde{\mathbf{X}}^t$ y dividiendo por $n-1$, para formar la matriz de varianza y covarianza, nos queda

$$\frac{1}{n-1}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}\,\mathbf{w} = \mathbf{S}\,\mathbf{w} = 0$$

Esta igualdad implica que la matriz $\mathbf{S}$ tiene un autovalor nulo y $\mathbf{w}$ es su autovector asociado, y además las coordenadas del vector $\mathbf{w}$ indican la combinación lineal entre las p variables.

Ejemplo 1. Se tiene la siguiente matriz de varianza y covarianza:

$$\mathbf{S} = \begin{pmatrix} 0.0947 & 0.0242 & 0.0054 & 0.0594 \\ 0.0242 & 0.0740 & 0.0285 & 0.0491 \\ 0.0054 & 0.0285 & 0.0838 & 0.0170 \\ 0.0594 & 0.0491 & 0.0170 & 0.0543 \end{pmatrix}$$

Calculando los autovalores de esta matriz mediante el software DERIVE nos arroja los siguientes valores

$$0.1729668172;\ 0.04616765904;\ 0.08761555169;\ 4.997202947 \times 10^{-5}$$

Puesto que este cálculo es una aproximación, según el algortimo del software, podemos sospechar que el cuarto autovalor de los anteriores es practicamente cero más aún si trabajamos hasta el cuarto dísgito significativo. Como sea el mismo programa nos arroja el siguiente autovector asociado al autovalor $4.997202947 \times 10^{-5}$(redondeando hasta el tercer dígito)

$$\begin{pmatrix} 0.408 & 0.408 & 0.000 & -0.816 \end{pmatrix}$$

Dividiendo cada componente de este autovector por la mayor componente, obtenemos el autovector

$$\begin{pmatrix} 0.5 & 0.5 & 0 & -1 \end{pmatrix}$$

Lo que nos estaría indicando que la cuarta variable resulta ser el promedio simple de las dos primeras.

El resultado anterior se puede generalizar a un mayor número de autovalores nulos. En efecto, si $\mathbf{S}$ tiene rango h ($h < p$) existirán $r = p - h$ combinaciones lineales entre las variables $\mathbf{X}$, y esto se concluye en virtud de que $\mathbf{S}$ y $\mathbf{X}$ (o $\widetilde{X}$) tienen el mismo rango[3].

4. Medidas globales de variabilidad

Interesa medir la variabilidad de un conjunto de variables, fundamentalmente si existe o no variación entre los atributos o preguntas que vamos a hacer. La propia matriz S nos puede entregar ciertas medidas de variabilidad, luego estas medidas las vamos a asocial al con-

[3] Nos basamos en el hecho de que $rango(A) = Rango(A^t) = Rango(A^tA) = Rango(AA^t)$.

cepto de distncia entre puntos 8donde los "puntos" serán las respuestas de cada individuo a nuestras preguntas).

4.1 La varianza total y la varianza media

Una primera forma de medir la variabilidad entre un conjunto de variables es mediante la suma de las varianzas de cada variable. Esto significa el cálculo de la traza de la matriz de varianzas y covarianzas. Se define entonces la *varianza total* de los datos como

$$T = traza(\mathbf{S}) = \sum_{i=1}^{p} s_i^2$$

y la *varianza media* por

$$\overline{s}^2 = \frac{1}{p} \sum_{i=1}^{p} s_i^2$$

El inconveniente de estas definiciones es que no considera una eventual estructura de dependencia entre las variables. En efecto, para un caso extremo consideremos $p = 2$, y de tal modo que de las dos variables que vamos a observar entre ellas se tenga la relación $\mathbf{y} = a + b\mathbf{x}$. En este caso, si $\mathbf{x}$ admite una varianza s_x^2, la varianza para $\mathbf{y}$ será $b^2 s_x^2$, de modo que la varianza total es $(1 + b^2)s_x^2$, y sin embargo es claro que la variabilidad connunta entre ambas variables (esto es, la covarianza) es nula. De otra forma, es posible que un connjunto de variables tenga una alta dependencia, y enconsecuencia la variabilidad conjunta será pequeña, no obstante la variabilidad total puede ser muy alta. Es decir, esta defimnición de variabilidad total no considera el grado de dependencia entre las variables.

4.2 La varianza generalizada

La *varianza generalizada* se define como el determinante de la matriz de varianza y covarianza, esto es

$$VG = \det \mathbf{S}$$

y la raíz cuadrada de VG se denomina *desviación típica generalizada*.

La interpretación que tiene esta varianza generalizada es bastante interesante. Suponga que $p = 2$, esto es tenemos dos variables, entonces la matriz $\mathbf{S}$ adopta la siguiente forma

$$S = \begin{pmatrix} s_x^2 & r\, s_x s_y \\ r\, s_x s_y & s_y^2 \end{pmatrix}$$

donde r es el coeficiente de correlación lineal de Pearson definido por $r = s_{xy}/s_x s_y$. El determinante de esta matriz es

$$VG = s_x^2 s_y^2 \left(1 - r^2\right)$$

sabemos que la interpretación de este determinante es el área. Notemos además que la desviación típica generalizada es

$$|\mathbf{S}|^{1/2} = s_x s_y \sqrt{(1 - r^2)} \tag{3}$$

Observemos que si las variables son independientes, entonces puesto que los valores observados de $\mathbf{x}$ están contenidos en un 90% en el intervalo de longitud $6s_x$, también los

valores observados de $\mathbf{y}$ estarán contenidos en un intervalo de longitud $6s_y$, entonces en virtud de la independencia entre ambas variables el 90 de los valores observados (x, y) estarán contenidos en un rectángulo de lados $6s_x$ y $6s_y$ [4]Y podemos ovbservar que el área de este rectángulo es $36s_x s_y$, de manera que el área coupada por estas variables es directamente proporcional a la desviación típica generalizada. Por otro lado, si las variables tienen una dependencia de tipo lineal entonces los valores observados conjuntamente tenderán a ubicarse cerca de la recta de regresión, y el área de contención será mucho menor, y analíticamente esto significa que $r^2 \to 1$, de modo que el áera ocupada por los datos tenderá hacia cero cuanto más aumente r^2. En el límite, cuando $r^2 = 1$, todos los puntos estarán en una línea recta cuya área es obviamente cero. De modo que la fórmula (3) describe esta contracción del área ocupada por las observaciones según aumente el coeficiente de correlación lineal.

4.3 La varianza efectiva

Se define la *varianza efectica* como

$$VE = (\det \mathbf{S})^{1/p}$$

Esta varianza es una generalización de la media geométrica, en efecto si $\mathbf{S}$ es una matriz diagonal, entonces la VE es la media geométrica de las varianzas de las variables. Se define la *desviación efectiva* como

$$DE = (\det \mathbf{S})^{1/2p}$$

5. Variabilidad y distancias

En lo que sigue supondremos conocida el signiificado de una distancia en $\mathbb{R}^p$ así como las proiedades más esenciales. En cualquier caso y en lo posible, haremos uso de las distancias sobre las filas (o columnas) de la matriz de datos $\mathbf{X}$, considerando a estas filas (o columnas) como puntos en el espacio dimensional real respectivo.

5.1 Distancia de Minkowski

Manteniendo prsente la matriz de datos

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

se define la distancia de Minkowski entre dos filas i y j de la matriz $\mathbf{X}$, como elementos de

[4] En efecto, por el teorema de Tchebyshev que dice que por lo menos el 90% de las observaciones estarán contenidas en el intervalo centrado en la media de longitud seis veces la desviación típica.

$\mathbb{R}^p$, de la siguiente forma:

$$d_{ij}^{(r)} = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^r \right)^{1/r}$$

distancia que depende del parámetro r. Los valores de r usados con más frecuencia son cuando $r = 2$ y $r = 1$. En el primer caso caemos en la llamada *distancia euclídea*, y en el segundo caso se llama distancia en L_1. La distancia más usada es la euclídea pero tiene el inconveniente que depende de las unidades de medidas de las variables. Supongamos que tenemos los datos para tres personas en que se ha medido su estatura, en metros, y su peso, en kilogramos:

persona	estatura (m)	peso (kgr)
1	1.80	80
2	1.70	72
3	1.65	81

la distancia euclídea entre la persona 1 y la persona 2 es

$$d_{12} = \sqrt{(1.80 - 1.70)^2 + (80 - 72)^2} = 8.0006$$

y la distancia entre las personas 1 y 3 es

$$d_{13} = \sqrt{(1.80 - 1.65)^2 + (80 - 81)^2} = 1.0112$$

De manera que podemos concluir que la persona 1 está más cerca de la persona 3 que de la persona 2. Supongamos ahora que realizamos un cambio de escala en la variable estatura, y trabajaremos con centímetros. La nueva tabla de datos es

persona	estatura (cm)	peso (kgr)
1	180	80
2	170	72
3	165	81

Nuevamente calculamos las distancias d_{12} y d_{13}

$$d_{12} = \sqrt{(180 - 170)^2 + (80 - 72)^2} = 12.806$$

$$d_{13} = \sqrt{(180 - 165)^2 + (80 - 81)^2} = 15.033$$

y como podemos abservar esta vez la persona 1 se encuentra más cerca de la persona 2 que de la 3. De modo que, con este ejemplo, podenos comprobar que la distancia euclíde depende mucho de las unidades de las variables, y no es aconsejable utilizarla en métodos multivariantes cuando no existe una unidad fija entre las variables.

Una manera de evitar este problema es dividir cada variable por un término que elimine el efecto de la escala. Esto nos conduce a una *distancia euclídea ponderada*, que se define como

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)^t \, \mathbf{M}^{-1} \, (\mathbf{x}_i - \mathbf{x}_j) \right]^{\frac{1}{2}} \tag{4}$$

donde $\mathbf{M}$ es una matriz diagonal que se utiliza para estandarizar las variables y hacer las medidas invariante ante cambios de escala. Si la matriz $\mathbf{M}$ es la matriz que en su diagonal

principal lleva las desciaciones típicas de las variables. la ecuación (4) queda como

$$d_{ij} = \left(\sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2 \right)^{\frac{1}{2}} = \left(\sum_{k=1}^{p} s_k^{-2} \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$$

que se interpreta como una distancia euclídea donde cada coordenada se pondera inversamente proporcional a la varianza.

En el ejemplo anterior, considerando la estatura y el peso en metros y kilogramos respectivamente, tenemos que la varianza de la estatura para las tres observaciones es de 58.333, y la varianza del peso para las tres observaciones es 24.333, de modo que las distancias ponderadas con estas varianzas son

$$d_{12} = \sqrt{\frac{1}{58.333}(180-170)^2 + \frac{1}{24.333}(80-72)^2} = 2.0842$$

$$d_{13} = \sqrt{\frac{1}{58.333}(180-165)^2 + \frac{1}{24.333}(80-81)^2} = 1.9744$$

Y podemos concluir que, con está métrica, que la persona 1 está más cerca de la persona 3 que de la 2.

En general, la matriz debe ser no singular y definida positiva (para que admita la definición de distancia). En el caso particular que $\mathbf{M} = \mathbf{I}$ se obtiene la distancia euclídea. En el caso en que $\mathbf{M} = \mathbf{S}$ se llega a la famosa distancia de Mahalanobis que estudiaremos a continuación.

5.2 La distancia de Mahalanobis

Se define la distancia entre dos "puntos" $\mathbf{x}_i$ y $\mathbf{x}_j$ de $\mathbb{R}^p$ como

$$d_{ij} = \left[\left(\mathbf{x}_i - \mathbf{x}_j \right)^t \mathbf{S}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j \right) \right]^{\frac{1}{2}}$$

Vamos a interpretar esta distancia y veremos que es una medida muy razonable de distancia entre variables correlacionadas. Como antes, consideremos el caso $p = 2$. Entonces, escribiendo $s_{12} = r\, s_1 s_2$, tenemos que el inverso de la matriz de varianza y covarianza es

$$\mathbf{S}^{-1} = \frac{1}{(1-r^2)} \begin{pmatrix} s_1^{-2} & -r\, s_1^{-1} s_2^{-1} \\ -r\, s_1^{-1} s_2^{-1} & s_2^{-2} \end{pmatrix}$$

y la distancia de Mahalanobis al cuadrado entre dos puntos (x_1, y_1), (x_2, y_2) es como sigue

$$d_{ij}^2 = \frac{1}{(1-r^2)} \left[\frac{(x_1-x_2)^2}{s_1^2} + \frac{(y_1-y_2)^2}{s_2^2} - 2r\, \frac{(x_1-x_2)(y_1-y_2)}{s_1 s_2} \right]$$

Si $r = 0$, esta distancia se reduce a la distancia euclídea estarandizando las variables por sus desviaciones típixcas. Cuando $r \neq 0$ la distancia de Mahalanobis añade un término que puede ser positivo, y, en este caso, "agrega" más distancia entre los puntos; o el término añadido puede ser negativo (y por lo tanto "junta" más los puntos). La cuestión es entonces interpretar el signo del término

$$-2r\, \frac{(x_1-x_2)(y_1-y_2)}{s_1 s_2}$$

Ete término es negativo si $r > 0$ y si las diferencias entre las variables tienen el mismo signo, o si $r < 0$ y si las diferencias entre las variables son de diferente signo. Por ejemplo, entre el peso y la estatura es natural pensar que hay correlación positiva, $r > 0$: al aumentar la estatura de una persona también lo hace su peso Luego si hay dos personas que cumplen esta correlación la distancia entre ellos será pequeña, que dos personas que no cumplan la correlación, es decir que una persona sea más alta que otra pero de menor peso que la misma, la distancia entre ellos será más grande. *"La capacidad de esta distancia para tener en cuenta la forma de un elemento a partir de su estructura de correlación, explica su introducción por P. C. Mahalanobis, en los años treinta del siglo pasado para comparar las medidas físicas de razas en la India"*.

6. Medidas de dependencia lineal

Un objetivo en el tratamiento de datos multivariantes es comprender la estructura de dependencia entre las variables. Como siempre desarrollaremos la teoría basándonos en la matriz de datos $\mathbf{X}$. Las dependencias pueden ocurrir en los siguientes sentidos

(a) entre par de variables, esto es si dos pares de columnas de la matriz $\mathbf{X}$ tienen algún grado de depndencia lineal,
(b) entre una variable y las demás, esto es si alguna columna de la matriz $\mathbf{X}$ puede depender linealmente de las otras $p - 1$ columnas,
(c) entre pares de variables pero eliminando el efecto de las demás, y
(d) entre el conjunto de todas las variables.

Vamos a analizar estos cuatro aspectos.

6.1 Dependencia por pares: la matriz de correlación

Este tipo de dependencia es la más utilizada en la estadística descriptiva, y su estudio se inicia con el cálculo y la interpretación (que suponemos conocida) del *coeficiente de correlación lineal de Pearson*. Supongamos que $\mathbf{x}_j$ y $\mathbf{x}_k$ son dos variables en $\mathbb{R}^p$, entonces este coeficiente es

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

y tiene las siguientes propiedades:

- $0 \leq r_{jk} \leq 1$;
- Si existe una relación exacta entre las variables, $x_{ij} = a + bx_{ik}, \ i = 1 \rightarrow p$, entonces $|rjk| = 1$;
- r_{jk} es invariante ante transformaciones lineales de las variables.

La dependencia por pares entre las variables, en nuesytro caso entre las columnas de la

matriz X, se mide simultáneamente por la matriz de correlación R definida como:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

Es una matriz cuadrada, simétrica, con unos sobre la diagonal principal, y evidentemente semidefinida positiva. Es sencillo de mostrar que su relación con la matriz de varianzas y covarianzas está dada por

$$\mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\ \mathbf{D}^{1/2}$$

donde $\mathbf{D}$ está definida como

$$D = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p^p \end{pmatrix}$$

la matriz diagonal de orden p formada por los elementos de la diagonal de $\mathbf{S}$. Y es de esta representación en que se deduce de que $\mathbf{R}$ es semidefinida positiva en cuanto y en tanto $\mathbf{S}$ lo es.

6.2 Dependencia de cada variable y el resto: regresión múltiple

Supongamos que por alguna razón estamos interesados en la variable $\mathbf{x}_j$, que para simplificar la notación denotaremos por $\mathbf{y}$ y llamaremos *variable predictora* (observe que $\mathbf{x}_j$ es una columna de nuestra matriz $\mathbf{X}$). Supongamos además que, a fortiori, queremos expresar la variable y como una combinación lineal de las variables restantes $\mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_p$ que llamaremos *variables explicativas o regresores*. Entonces, nuestra intención es encontrar los "mejores" valores de $\widehat{\beta}_k$ para encontrar

$$\widehat{y}_i = \overline{y} + \widehat{\beta}_1(x_{i1} - \overline{x}_1) + \cdots + \widehat{\beta}_p(x_{ip} - \overline{x}_p)\,; \quad i = 1,\, ...,\, n$$

de tal forma que el siguiente valor sea mínimo

$$M = \sum_{i=1}^{n} e_i^2$$

donde

$$e_i = (y_i - \widehat{y}_i)$$

La obtención de los valores $\widehat{\beta}_k$ bajo la condición de hacer mínimo $M = \sum_{i=1}^{n} e_i^2$ se realiza por el método de los multiplicadores de Lagrange, de modo que si llamamos al vector de los parámetros por $\widehat{\boldsymbol{\beta}}$ la solución es

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}_R^t \mathbf{X}_R\right)^{-1} \mathbf{X}_R^t \mathbf{y}$$

donde $\mathbf{X}_R$ es la matrz de $n \times (p-1)$ que se obtiene de la matriz centrada de datos $\widetilde{\mathbf{X}}$ que se obtiene al eliminar la columna que corresponde a la variable que queremos prever $\mathbf{y}$.

El promedio corregido de los residuos al cuadrado, o varianza, de esta ecuación de

regresión múltiple para explicar $\mathbf{x}_j$ es

$$s_r^2(j) = \frac{\sum e_i^2}{n-1} \tag{5}$$

y es una medida de precisión de la regresión. Se obtiene una medida adimensional de la dependencia partiendo de la identidad

$$y_i - \overline{y} = \widehat{y}_i - \overline{y} + e_i$$

elevando al cuadrado y sumando se verifica fácilmente que

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} e_i^2$$

donde $VT = \sum_{i=1}^{n} (y_i - \overline{y})^2$ se expresa como la *variabilidad total* de los datos (observe que solo falta dividir por $n-1$ para obtener la varianza de $\mathbf{y}$); $VE = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$ es la *variabilidad explicada* por la regresión; y $VNE = \sum_{i=1}^{n} e_i^2$ es la variabilidad no explicada o *residual*. Una medida descriptiva de la capacidad predictiva del modelo es el cocientre entre la variabilidad explicada por la regresión y la variabilidad total, y tal medida se llama *coeficiente de determinación* o coeficiente de *correlación múltiple* al cuadrado, y se denota por

$$R^2_{j\cdot 1,\ldots,p} = \frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

Por la ecuación (1), podemos escribir

$$R^2_{j\cdot 1,\ldots,p} = 1 - \frac{s_r^2(j)}{s_j^2}$$

No lo vamos a demostrar en estos apuntes pero existe un algoritmo para calcular los valores $s_r^2(j)$ a partir de la matriz de varianza y covarianza $\mathbf{S}$. Y es el siguiente:

(1) Invierta la matriz $\mathbf{S}$, y tome el elemento j-ésimo de la diagonal de $\mathbf{S}^{-1}$,
(2) llame a este elemento selecionado s^{jj}, entonces $s^{jj} = 1/s_r^2(j)$

De manera que con el algoritmo anterior podemos calcular, mediante la matriz $\mathbf{S}$, todos los coeficientes de correlación multiple, esto es

$$R^2_{j\cdot 1,\ldots,p} = 1 - \frac{1}{s^{jj} s_{jj}} \quad ; \quad j = 1 \to p$$

entendiendo que $s_{jj} = s_j^2$ el elemento j-ésimo de la diagonal de $\mathbf{S}$.

Como podemos observar, podemos obtener todos los coeficientes de correlación múltiple entre una variable y las restantes a partir ede las matrices $\mathbf{S}$ y $\mathbf{S}^{-1}$.