Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz

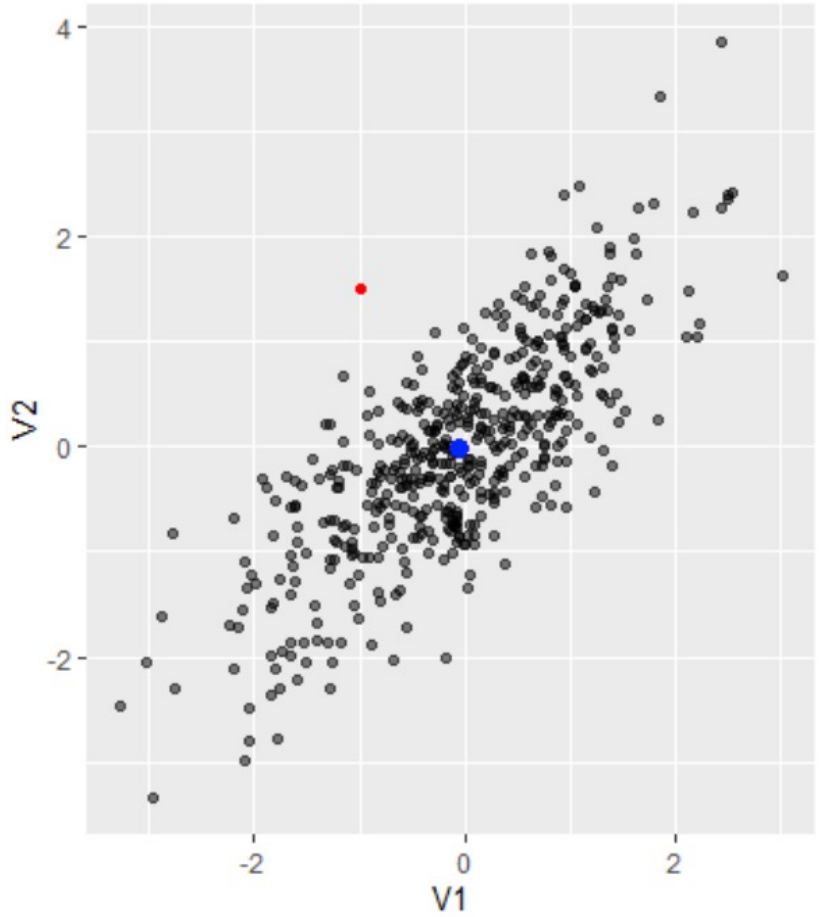
Octubre, 2022

Distancia de Mahalanobis (1)

- Percepciones iniciales…
  - Uno de los objetivos principales y difíciles en estadística, es determinar cuándo dos elementos son parecidos y cuándo no.
  - Resolver adecuadamente esta cuestión es la base de diversos procedimientos y algoritmos, desde la generación de hipótesis hasta métodos de clasificación.
  - Entonces para ello, podemos usar la palabra ambigua *elementos*, porque a veces nos interesa comparar magnitudes escalares (números), pero otras veces queremos comparar observaciones multivariantes (vectores).
  - Incluso podemos necesitar saber si dos funciones se parecen, o elementos más complicados, como es el caso de imágenes, fotografías, documentos, tweets, fragmentos de ADN, eventos delictivos, entre otros tienen cierta similitud.

Distancia de Mahalanobis (2)

- Para introducir la idea básica se consideran los datos siguientes…
  - A simple vista, parece claro que el punto representado en color rojo resulta atípico y está "lejos" de la media (azul). Esto no se debe a que el valor de cada una de las variables por separado sea extraño.
  - De hecho, en el punto rojo la primera variable (V1) vale -1 y hay muchos otros puntos para los que este valor es menor o mayor. Lo mismo ocurre para la segunda variable (V2), que vale 1.5.
  - El punto rojo es atípico porque en su caso el valor de V2 es atípico, comparado con el de los puntos en los que V1 es aproximadamente -1.
  - La relación entre las variables en este punto es ligeramente distinta que en el resto, por eso destaca.

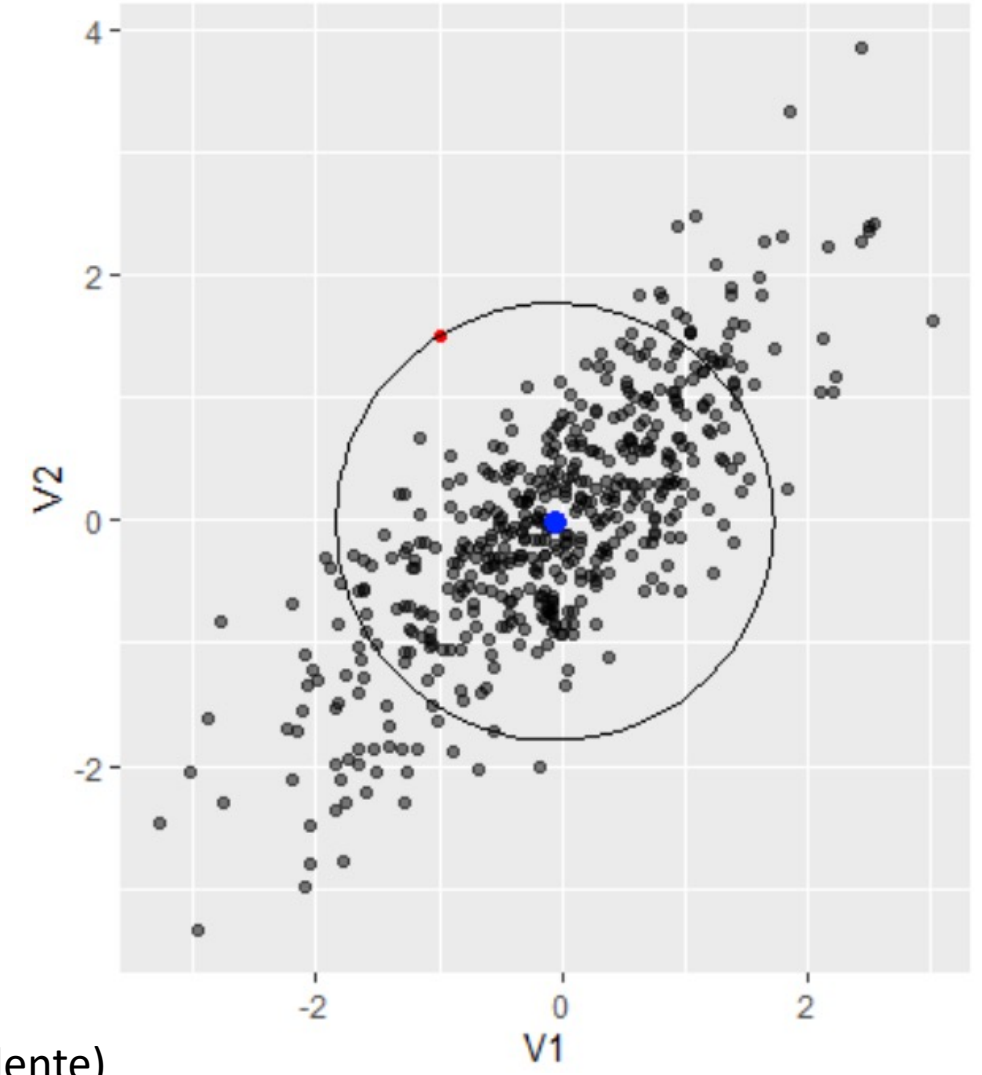


Distancia de Mahalanobis (3)

- Si usamos la distancia Euclídea habitual para medir la distancia a la media no tendremos en cuenta la relación entre las variables.
- Muchos puntos que estadísticamente aparecen como más "normales" (y están fuera del círculo) distarán de la media más que el punto rojo, cuando estadísticamente destacan menos.
- Para definir una distancia que tenga sentido estadístico tenemos que tener en cuenta las correlaciones entre las variables.
- Entonces si x es un vector aleatorio (de dimensión p), procedente de una distribución con media μ y matriz de covarianzas $\sum$, la distancia de Mahalanobis entre x y μ es:

$$d_M(X,\mu) = \sqrt{(X-\mu)^T \sum^{-1} (X-\mu)}$$

$$d_M^2(X,\mu) = (X-\mu)^T \sum^{-1} (X-\mu)$$ (Forma equivalente)



Distancia de Mahalanobis (4)

- En la práctica no se conoce ni el vector de medias ni la matriz de covarianzas poblacionales, por lo que se reemplazan por sus equivalentes muestrales, y se usa la fórmula:

$$d_M(x, \bar{x}) = \sqrt{(x - \bar{x})^T \sum^{-1} (x - \bar{x})}$$

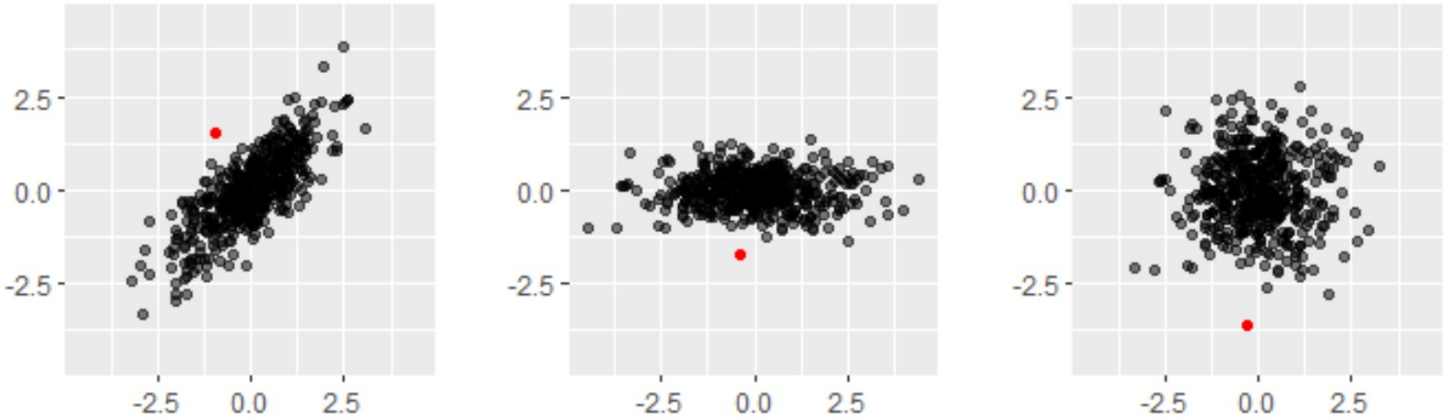
- En el caso de tener observaciones escalares en lugar de vectores, si $\bar{x}$ es la media muestral y s es la desviación típica muestral, la expresión anterior es simplemente la observación estandarizada en valor absoluto:

$$d_M(x, \bar{x}) = \sqrt{\frac{(x, \bar{x})^2}{s^2}} = \frac{|x - \bar{x}|}{s}$$

- Una buena forma de entender la definición para vectores es a través de una sucesión de operaciones que pueden hacerse sobre los datos. El cálculo de la distancia de Mahalanobis equivale a:
  1. Trasladar los datos de forma que su nueva media sea el origen de coordenadas.
  2. Rotarlos para que las correlaciones entre las variables sean todas iguales a cero. (lo que no cambia el resultado final).
  3. Estandarizar el resultado del paso anterior para que todas las varianzas valgan uno. Así se consigue una distancia que no depende de las unidades de medida, o de si las magnitudes de las variables son muy diferentes entre sí.
  4. Calcular la distancia Euclídea al origen de los puntos resultantes.

Distancia de Mahalanobis (5)

- En la figura se representa el resultado de los tres primeros pasos con los datos del ejemplo.
  - Se ha marcado en rojo el punto que habíamos identificado como atípico en un principio.
  - Se puede ahora comprobar que, ese punto rojo en el sentido de la distancia de Mahalanobis, es uno de los puntos más lejanos a la media, lo que coincide con la intuición que se tenía al mirar los datos.
- En la figura se representan los puntos en distintos tonos según su distancia de Mahalanobis a la media.
  - Cuanto más claro, más cercano.
  - Las curvas de nivel de la distancia son elipses.
  - La dirección de los ejes está determinada por la correlación entre las variables.



Distancia de Mahalanobis (6)

- Definición
  - Es una función que permite determinar la distancia entre dos individuos definidos por *p* variables, en donde se consideran la matriz de covarianzas de las variables que definen a los componentes de cada individuo.
- Características
  - Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales.
  - Se utiliza con variables aleatorias que tienen la misma distribución de probabilidad.
  - Es la forma más utilizada para medir la distancia entre vectores $d_M(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \sum^{-1} (\vec{x}-\vec{y})}$
  - La diferencia con la distancia Euclídea es que la distancia Mahalanobis pondera respecto a la varianza, dándole un mayor peso a aquellos atributos con varianza menor que aquellos con varianza mayor.
  - Debe cumplir con 3 propiedades para ser una distancia: (1) semipositividad, (2) simetría y (3) desigualdad triangular.

$$d(a,b) \geq 0 \; \forall a,b \in X \wedge \; d(b,a) = 0 \; si \; a = b$$
$$d(a,b) = d(b,a) \; \forall a,b \in X$$
$$d(a,b) \leq d(a,c) + d(c,b) \; \forall a,b,c \in X$$

Distancia de Mahalanobis (7)

- Características:
  - Para usar Mahalanobis para clasificar un punto de prueba como perteneciente a una de las *N* clases, primero se estima la matriz de covarianza de cada clase, generalmente basándose en muestras que se sabe pertenecen a cada clase.
  - Luego, dada una muestra de prueba, se calcula la distancia de Mahalanobis a cada clase y se clasifica el punto de prueba como perteneciente a esa clase para la cual la distancia de Mahalanobis es mínima.
- Aplicaciones más comunes:
  - Detección de valores atípicos en modelos de regresión lineal.
    - Se dice que un punto que tiene una mayor distancia de Mahalanobis del resto de la población de puntos de la muestra, tiene un mayor apalancamiento, ya que tiene una mayor influencia en la pendiente o los coeficientes de la ecuación de regresión.
  - Construcción de clusters y análisis de conglomerados.
    - Ampliamente utilizada en técnicas de clasificación. Está estrechamente relacionada con la distribución T-cuadrada de Hotelling que se utiliza para las pruebas estadísticas multivariadas y el análisis discriminante lineal de Fisher enfocado a clasificación supervisada.
  - Detección valores atípicos multivariantes.
    - Utilizando técnicas de regresión para determinar si un caso específico dentro de una población de muestra es un valor atípico, mediante la combinación de dos o más puntuaciones variables. Incluso para distribuciones normales, un punto puede ser un valor atípico multivariado, sino es un valor atípico univariante para ninguna variable (i.e., densidad de probabilidad concentrada a lo largo de la línea), lo que hace que la distancia de Mahalanobis sea una medida más sensible, que verificar las dimensiones individualmente.

Distancia de Mahalanobis (8)

Variables

$$X = \begin{bmatrix} x_1 & x_2 & & x_j & & x_p \\ x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ n \end{matrix}$$

Individuos

x_{ij} = **Al valor que toma la variable en el** $i-$ ***ésimo*** **individuo**

- Finalmente $c = \begin{pmatrix} \overline{x_1} & \overline{x_2} & \cdots & \overline{x_j} & \cdots & \overline{x_p} \end{pmatrix}$ es el centroide de la matriz de datos, es decir, es un individuo hipotético, cuyas componentes están formadas por las medias aritméticas de las variables.

- Su cálculo consiste en la definición de una matriz de datos *X*, que es una matriz numérica de orden $n \times p$ que recopila información p que es una variante asociada con n individuos.
- En la figura se muestra la forma que presenta la matriz de datos, en donde:
  - Las columnas son los valores de las variables, y
  - Las filas son las componentes de cada *n* individuo que participa.
  - Dicho de otra forma x_j es el vector columna (vector atributo) que contiene los valores correspondientes a la variable x_j

$$x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix}$$

- $x^{(i)}$ es el vector fila (vector individuo) que contiene los valores correspondientes al $\boldsymbol{i}$**–ésimo** individuo.

$$x^{(i)} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \end{pmatrix}$$

Distancia de Mahalanobis (9)

- Por tanto, la distancia de Mahalanobis:
  - Sea $M_{n\times p}$ una matriz de datos, en donde la distancia cuadrática entre los individuos i y j se denota por $d_M(i,j)$ que es el valor resultante de aplicar el producto matricial:

$$d_M^2(i,j) = \left(x^{(i)} - x^{(j)}\right)\sum^{-1}\left(x^{(i)} - x^{(j)}\right)^T$$

NOTA. Si quisiéramos la distancia hay que sacar la $\sqrt{(i,j)}$

  - Siendo $\sum$ la matriz de covarianzas muestrales $\sum_{ij} = cov(x_i, x_j), i = 1, \dots, p;\ j = 1, \dots, p$, siempre y cuando dicha matriz sea no singular (matriz inversa).
  - La distancia cuadrática de Mahalanobis del *i*-ésimo individuo al centroide de la matriz de datos, se designa como $d_M^2(i)$.

$$d_M^2(i) = \left(x^{(i)} - c\right)\sum^{-1}\left(x^{(i)} - c\right)^T$$

  - $d_M^2(i)$ permite calcular la distancia cuadrática de Mahalobis entre cada individuo y el centroide por separado; sin embargo, la diagonal de la matriz se define como sigue:

$$diag = \left(I_n - \frac{1}{n}J_n\right)X\sum^{-1}X^t\left(I_n - \frac{1}{n}J_n\right), \quad I_n = \text{matriz de identidad } n\times n,\ J_n = \text{matriz } n\times n \text{ de unos}$$

  - Esta diagonal de la matriz proporciona todas las distancias de cada individuo al centroide.

Distancia de Mahalanobis (10)

- Ejemplo, considérese la matriz de datos formada por 6 individuos y 4 variables:

$$X = \begin{bmatrix} 2 & 4 & 1 & 7 \\ 1 & 6 & 3 & 5 \\ 1 & 8 & 4 & 2 \\ 4 & 5 & 4 & 0 \\ 5 & 0 & 6 & 1 \\ 6 & 1 & 8 & 2 \end{bmatrix}$$

- Los 6 individuos están representados por los siguientes 6 vectores fila:

$$x^{(1)} = (2 \quad 4 \quad 1 \quad 7) \qquad x^{(2)} = (1 \quad 6 \quad 3 \quad 5)$$

$$x^{(3)} = (1 \quad 8 \quad 4 \quad 2) \qquad x^{(4)} = (4 \quad 5 \quad 4 \quad 0)$$

$$x^{(5)} = (5 \quad 0 \quad 6 \quad 1) \qquad x^{(6)} = (6 \quad 1 \quad 8 \quad 2)$$

Distancia de Mahalanobis (11)

- Las 4 variables están definidas por los 4 vectores columnas:

$$x_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 4 \\ 5 \\ 6 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 4 \\ 6 \\ 8 \\ 5 \\ 0 \\ 1 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \\ 6 \\ 8 \end{bmatrix} \qquad x_4 = \begin{bmatrix} 7 \\ 5 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

- La matriz de covarianzas muestrales $\sum$ o $\boldsymbol{S}$ viene dada por la siguiente expresión:

$$\sum = \frac{1}{6-1} \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix}$$

Se dice matriz de cuasicovarianzas porque el denominador en lugar de *n* es *n-1*

$$s_{ij} = \sum_{k=1}^{6} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$S = \begin{bmatrix} 4.5677 & -5.6000 & 4.1333 & -3.1667 \\ -5.6000 & 9.2000 & -4.4000 & 1.6000 \\ 4.1333 & -4.4000 & 5.8667 & -4.3333 \\ -3.1667 & 1.6000 & -4.3333 & 6.9667 \end{bmatrix}$$

Distancia de Mahalanobis (12)

- El centroide de la matriz de datos es el siguiente vector fila, formado por las medias aritméticas de cada variable:

$$c = (\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_3 \quad \bar{x}_4) = (3.1667 \quad 4.0000 \quad 4.3333 \quad 2.833)$$
$$\mu = (\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_3 \quad \bar{x}_4) = (3.1667 \quad 4.0000 \quad 4.3333 \quad 2.833)$$

- La distancia cuadrática de Mahalanobis entre los individuos 1 y 3 es por definición la siguiente expresión:

$$\left(x^{(1)} - x^{(3)}\right)\Sigma^{-1}\left(x^{(1)} - x^{(3)}\right)^T$$

$$\left(x^{(1)} - x^{(3)}\right) = (2 \quad 4 \quad 1 \quad 7) - (1 \quad 8 \quad 4 \quad 2)$$

$$= (1 \quad -4 \quad -3 \quad 5)$$

$$d_M^2\left(x^{(1)}, x^{(3)}\right) = (1 \quad -4 \quad -3 \quad 5)\begin{bmatrix} 4.5677 & -5.6000 & 4.1333 & -3.1667 \\ -5.6000 & 9.2000 & -4.4000 & 1.6000 \\ 4.1333 & -4.4000 & 5.8667 & -4.3333 \\ -3.1667 & 1.6000 & -4.3333 & 6.9667 \end{bmatrix}^{-1} (1 \quad -4 \quad -3 \quad 5)^T = 9.7919$$

Distancia de Mahalanobis (13)

- Ahora la distancia cuadrática de Mahalanobis entre el individuo 4 y el centroide es:

$$\left(x^{(4)}-c\right)\Sigma^{-1}\left(x^{(4)}-c\right)^{T}$$

$$\left(x^{(4)}-c\right)=(4\quad 5\quad 4\quad 0)-(3.1667\quad 4.0000\quad 4.3333\quad 2.8333)$$

$$=(0.8333\quad 1.0000\quad -0.3333\quad -2.8333)$$

- Por lo tanto, aplicando el concepto, se tiene que esa distancia del individuo 4 al centroide es:

$$d_M^2\left(x^{(4)},c\right)=(0.8333\quad 1.000\quad -0.3333\quad -2.8333)\begin{bmatrix}4.5677 & -5.6000 & 4.1333 & -3.1667\\ -5.6000 & 9.2000 & -4.4000 & 1.6000\\ 4.1333 & -4.4000 & 5.8667 & -4.3333\\ -3.1667 & 1.6000 & -4.3333 & 6.9667\end{bmatrix}^{-1}(0.8333\quad 1.000\quad -0.3333\quad -2.8333)^{T}=3.77667$$

- Ahora, vamos a determinar todas las distancias cuadráticas de Mahalanobis de cada uno de los individuos al centroide.

Distancia de Mahalanobis (14)

- Entonces se define la matriz de densidad de orden 6, junto con la matriz de unos de orden 6, la matriz de datos *X* y la matriz de cuasicovarianzas muestrales:

$$I_6 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad J_6 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad X = \begin{bmatrix} 2 & 4 & 1 & 7 \\ 1 & 6 & 3 & 5 \\ 1 & 8 & 4 & 2 \\ 4 & 5 & 4 & 0 \\ 5 & 0 & 6 & 1 \\ 6 & 1 & 8 & 2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 4.5677 & -5.6000 & 4.1333 & -3.1667 \\ -5.6000 & 9.2000 & -4.4000 & 1.6000 \\ 4.1333 & -4.4000 & 5.8667 & -4.3333 \\ -3.1667 & 1.6000 & -4.3333 & 6.9667 \end{bmatrix}$$

Distancia de Mahalanobis (15)

- Por lo tanto, aplicando la fórmula de la diagonal, se obtiene el vector columna que representa las distancias cuadráticas de Mahalanobis de cada individuo al centroide:

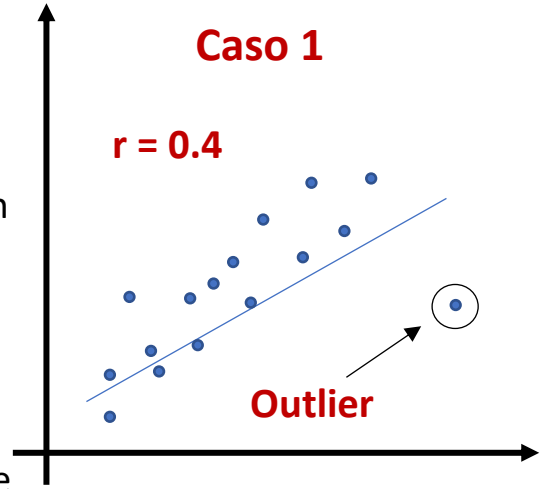
$$diag = \left(I_n - \frac{1}{n}J_n\right) X S^{-1} X^t \left(I_n - \frac{1}{n}J_n\right) = \begin{bmatrix} 3.605 \\ 1.618 \\ 2.713 \\ 3.776 \\ 4.124 \\ 4.163 \end{bmatrix}$$

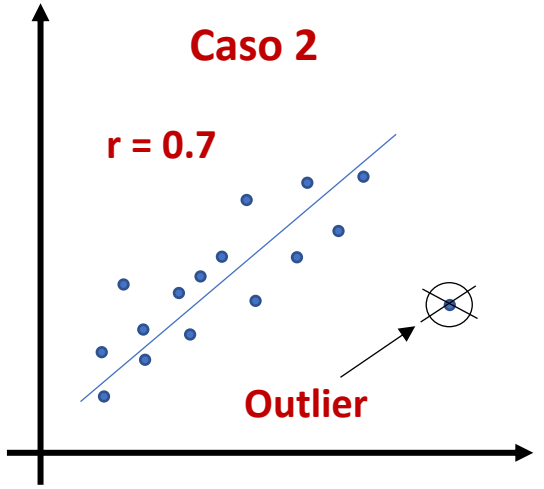
$$d_M^2\left(x^{(1)}, c\right) = 3.605$$
$$d_M^2\left(x^{(2)}, c\right) = 1.618$$
$$d_M^2\left(x^{(3)}, c\right) = 2.713$$
$$d_M^2\left(x^{(4)}, c\right) = 3.776$$
$$d_M^2\left(x^{(5)}, c\right) = 4.124$$
$$d_M^2\left(x^{(6)}, c\right) = 4.163$$

Distancia de Mahalanobis (16)

Conclusiones:

- Se recupera la idea de *correlación lineal bivariada*, con un problema específico:
  - *Casos atípicos para una distribución de pares ordenados*.
- Las distancias de los puntos respecto a una recta de regresión contribuyen a la magnitud del coeficiente.
- Existe un problema cuando solo algunos valores atípicos se apartan del conjunto general de pares ordenados (*x,y*) que ubican la recta de regresión casi paralela al eje de la abscisa.
- En tal caso el coeficiente de correlación se aproxima a cero y se interpreta como independencia entre las variables.
- Solo la inspección de la nube de puntos puede determinar si efectivamente las variables son independientes o su correlación sufre la influencia de algunos valores atípicos (*outliers*).
- Si observamos las gráficas considerando las diferencias en el coeficiente de correlación cuando se calcula dicho coeficiente, con la presencia del *outlier*, o cuando éste es removido del conjunto de datos.
- En el caso 1, se aprecia que la presencia del caso atípico, da como resultado una correlación de Pearson $r = 0.4$; si se remueve este outlier, dicha correlación aumenta en su magnitud $r = 0.7$.
- Los casos atípicos son observaciones infrecuentes, pero tienen una gran influencia en la pendiente de regresión puesto que ésta se basa en minimizar la suma de los cuadrados de las distancias de cada punto de la nube (par ordenado) a la línea teórica.
- Debido al efecto de un solo caso atípico la pendiente puede variar y el coeficiente puede cambiar drásticamente.

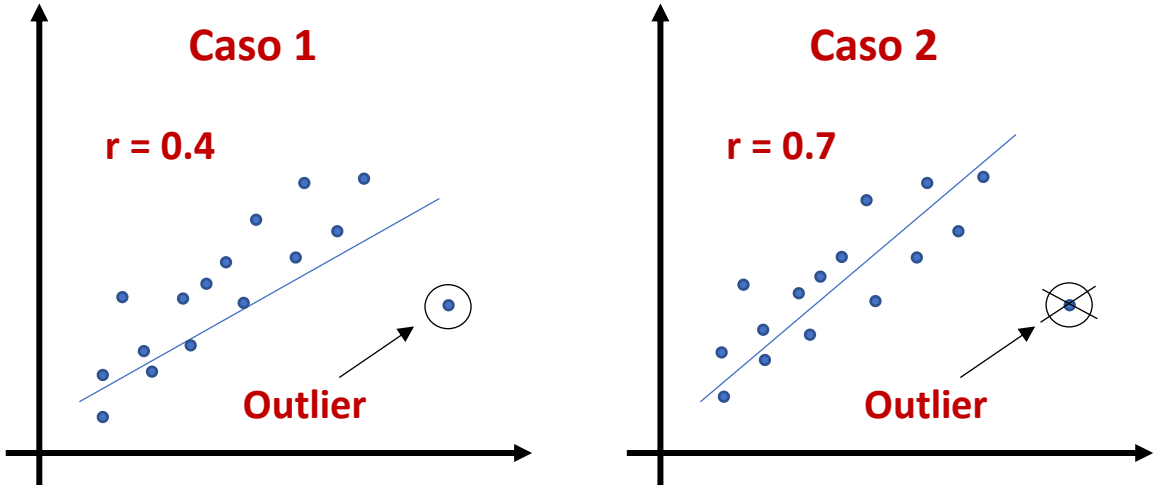




Un **outlier** es una observación anormal y extrema en una muestra estadística o serie temporal de datos que puede afectar potencialmente a la estimación de los parámetros del mismo

Distancia de Mahalanobis (17)

- Conclusiones
  - No se puede sugerir que los casos atípicos sean eliminados sin un análisis, puesto que no siempre son errores de medición.
  - Existe el problema de que uno o varios casos atípicos se hayan registrado en una muestra particular, pero que sucesivas muestras pueden determinar que éstos son en verdad valores ordinarios.
  - Sin embargo, muchas veces solo contamos con una sola muestra y determinar la naturaleza del valor atípico observado no resulta posible. En estas situaciones, un cálculo de las distancias entre el conjunto de valores es una opción recomendable, para manejar datos con casos atípicos.



Proyecto 3 (1)

Proyecto 3. Programa en R que permita calcular la distancia de Mahalanobis sobre el conjunto de datos Iris, para las 3 clases de flores que están descritas en el conjunto de datos.

1. **Cuantificar el valor de la distancia de Mahalanobis entre las flores: Setosa, Versicolor y Virginica, tomando como base la longitud y anchura del sépalo y pétalo de cada flor.**
2. **Graficar los valores de las distancias calculadas.**

Proyecto 4 (1)

Proyecto 4. Considerar los datos económicos de la siguiente tabla (dados en millones de dólares) de corporaciones industriales

1. **Calcular la distancia de Mahalanobis entre Ford y Exxon**
2. **Calcular la distancia de Mahalanobis entre General Motors e IBM**
3. **Calcular la distancia de Mahalanobis entre Philip Morris y Texaco**

Compañía	Ventas	Beneficios	Bienes
General Motors	126,974,000	4,224,000	173,297,000
Ford	96,933,000	3,835,000	160,893,000
Exxon	86,656,000	3,510,000	83,219,000
IBM	63,438,000	3,758,000	77,734,000
General Electric	55,264,000	3,939,000	128,344,000
Mobil	50,976,000	1,809,000	39,080,000
Philip Morris	39,069,000	2,946,000	38,528,000
Chrysler	36,156,000	359,000	51,038,000
DuPont	35,209,000	2,480,000	34,715,000
Texaco	32,416,000	2,413,000	25,636,000