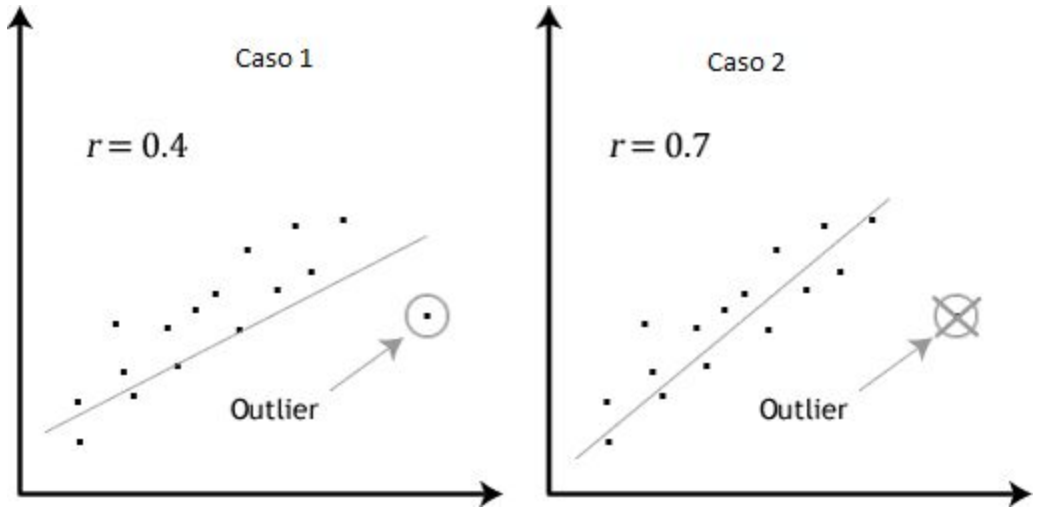
Distancia de Mahalanobis

Programa de la unidad didáctica de Estadística Básica
Prof. Jorge Lorenzo

VISIÓN GENERAL Y OBJETIVOS

En esta lección se recupera la idea de correlación lineal bivariada, tratando un problema específico: los casos atípicos para una distribución de pares ordenados. Según se mostró en el tutorial de estadística[1], 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 independiente o su correlación sufre la influencia de algunos valores atípicos (outliers). Veamos las siguientes gráficas considerando las diferencias en el coeficiente de correlación cuando se calcula dicho coeficiente con la presencia del caso atípico (outlier), o cuando este es removido del conjunto de datos.



En el caso 1, se aprecia que la presencia del caso atípico, da como resultado una

[1] El volumen Tutoriales de Estadística puede recuperarse del siguiente enlace https://ansenuza.unc.edu.ar/comunidades/handle/11086.1/1202

correlación r de Pearson igual a: r=0.4; si se remueve este caso atípico, dicha correlación aumenta en su magnitud a: 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. Consecuentemente, debido al efecto de un solo caso atípico la pendiente puede variar y el coeficiente puede cambiar drásticamente.

No se puede sugerir que los casos atípicos sean eliminados sin un examen de los mismo, 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 determinen 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 situaciones como la que estamos planteando un examen de las distancias entre el conjunto de valores es una opción recomendable. En esta lección revisaremos el concepto de distancia de Mahalanobis con una aplicación a datos con casos atípicos.

Distancias de Mahalanobis

La distancia de Mahalanobis es una medida de distancia alternativa a las distancias euclídeas. Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales, teniendo en cuenta la correlación entre ellas. La distancia de Mahalanobis entre dos variables aleatorias con la misma distribución de probabilidad se define formalmente como:

$$d_m = (\vec{x}; \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}$$

La distancia de Mahalanobis debe cumplir con tres propiedades, necesarias para ser una distancia, estas son: a) semipositividad, b) simetría y, c) desigualdad triangular. No nos extenderemos sobre estos particulares, pero se recomienda consultar tales propiedades en bibliografía especializada[2].

Con un ejemplo práctico se puede apreciar la utilidad de la distancia de Mahalanobis se.

[2] Salas Plata, Jorge & Portillo, María. (2008). P. Ch. Mahalanobis y las aplicaciones de su distancia estadística. CULCyT: Cultura Científica y Tecnológica, ISSN 2007-0411, Nº. 27, 2008, pags. 13-20. 5.

Supongamos que un horticultor desea clasificar dos tipos de manzanas: a) estándar y b): premium. Cada manzana se mide por su diámetro y su peso. Con estos datos se construye un vector para cada una de las manzanas que han de ser clasificadas.

$$\vec{x}_i \left(x_1; x_2 \right)^T$$

El diámetro y el peso de las manzanas no tienen varianzas iguales, por caso, se puede esperar mayor variación en los diámetros que en el peso. Si se calculan las distancias euclídeas, se estaría dando más importancia a la variable con menor varianza. El método de cálculo de distancias de Mahalanobis sirve para igualar la importancia de ambas variables en el resultado final. La expresión quedaría:

$$d_2\left(\vec{x}_1; \vec{x}_2 \right) = \sqrt{\frac{\left(x_{11} - x_{12} \right)^2}{\sigma_1} + \frac{\left(x_{21} - x_{22} \right)^2}{\sigma_2}}$$

Donde σ_i es la desviación estándar de la componente i de los vectores de medidas. Una propiedad de las distancias de Mahalanobis es que en la matriz vectorial se puede incluir la matriz de covarianza que da cuenta de la correlación entre las variables. Para nuestro ejemplo, el diámetro y el peso de las manzanas claramente están correlacionadas y tienen que ser consideradas juntas.

OBJETIVOS

1. Graficar una distribución de variables x e y.
2. Identificar casos atípicos.
3. Utilizar el procedimiento distancias de Mahalanobis.

MATERIAL NECESARIO

Para realizar esta actividad es necesario contar con una computadora con el paquete estadístico R instalado. En caso de no contar con este programa se puede recurrir a la página del Proyecto R, allí encontrará toda la información necesaria para instalar los paquetes y sus dependencias.

https://www.r-project.org/

ACTIVIDAD

Vamos a presentar un ejemplo de cómo calcular las distancias de Gauss y Mahalanobis en el paquete estadístico R. Lo primero que debemos hacer es generar el conjunto de datos necesario, vamos a generar dos variables a las que llamaremos prueba1 y prueba2. Para la carga de datos seguimos el siguiente script.

```
prueba1= c( 154, 157, 158, 160, 161, 160, 161, 162, 162, 164, 162, 162, 164, 156, 166, 170)

prueba2= c(63, 66, 68, 69, 68, 71, 70, 70, 71, 72, 72, 73, 73, 75, 76, 78)
```

Luego utilizamos la función data.frame() para crear un juego de datos en R

```
datos <- data.frame(prueba1 ,prueba2)
```

El método de distancia de Gauss normaliza todas las variables bajo una misma escala [función en R: scale].

Para ello debemos generar un vector boleano indicando los valores que esten a una distancia de más de 2 desviaciones estándar de la media.

```
prueba1.outlier <- abs(scale(datos$prueba1)) > 2

prueba2.outlier <- abs(scale(datos$prueba2)) > 2
```

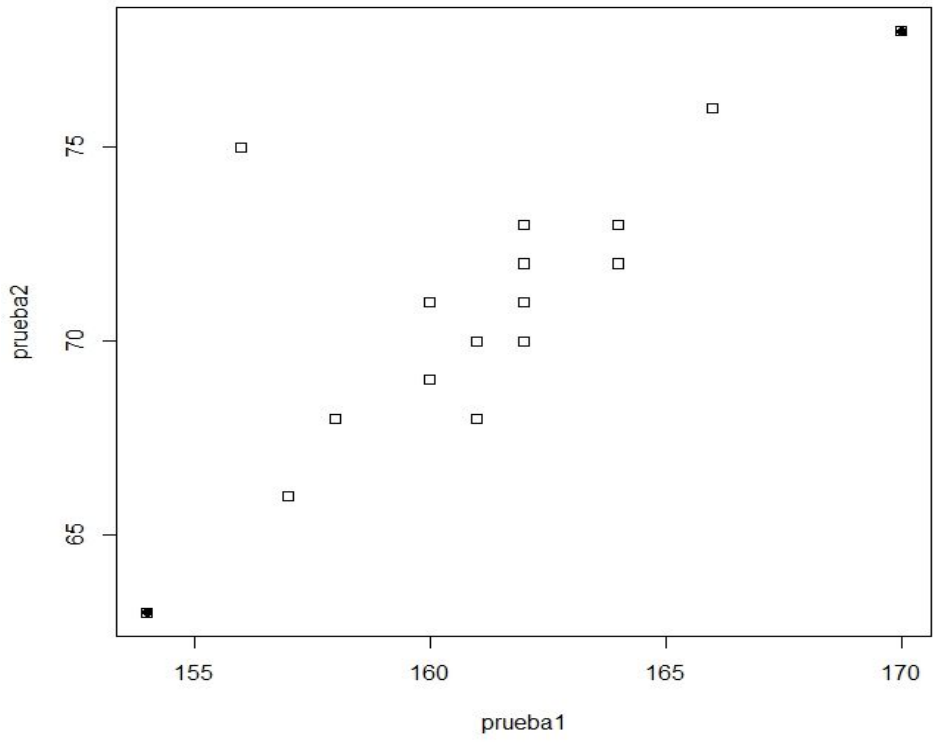
Se necesita almacenar los outlier encontrados para poder mostrarlos gráficamente

```
outlier <- rbind(datos[prueba1.outlier ,], datos[prueba2.outlier ,])
```

Para visualizar el gráfico con los datos destacando sus outlier

```
plot(datos, pch=0)
points(outlier , pch=16)
```

El gráfico generado por el programa es el siguiente:



Como mencionamos al principio, el método de distancia Mahalanobis mejora el método

de cálculo de distancia de Gauss al considerar la correlación entre las variables a analizar. Para su cálculo debemos emplear [función en R: mahalanobis]. Para el ejemplo que estamos analizando:

1) Determinar el número de outlier que queremos encontrar.

```
num.outliers <- 2
```

2) Ordenar los datos de mayor a menor distancia, según la métrica de Mahalanobis.

```
mah.ordenacion <- order(mahalanobis(datos , colMeans( datos), cov(datos)), decreasing=TRUE)
```

3) Generar un vector boleano los dos valores más alejados según la distancia Mahalanobis.

```
outlier2 <- rep(FALSE , nrow(datos))

outlier2[mah.ordenacion[1:num.outliers]] <- TRUE
```

4) Resaltar con un punto relleno los 2 valores outliers.

```
colorear.outlier <- outlier2 * 16
```

5) Visualizar el gráfico con los datos destacando sus outlier.

```
plot(datos , pch=0)

points(datos , pch=colorear.outlier)
```

El gráfico que se muestra a continuación contiene los dos casos atípicos calculados con el método de distancias de Mahalanobis.