



Probabilidad y Estadística - Grado en Matemáticas
Curso 2021-2022


Práctica de ordenador del Tema 2

Actividad 1: En relación con el Ejemplo 1 de los apuntes (datos de “familias.txt”), elaborar un script de  que permita realizar las tareas descritas a continuación.

- (a) Leer los datos del fichero “familias.txt”.
- (b) Obtener el vector de medias muestral.
- (c) Obtener la matriz de covarianzas muestral, la matriz de covarianzas entre las mediciones del primero y del segundo hijos, y la matriz de correlaciones.
- (d) ¿Cómo se podría obtener la matriz de correlaciones, a partir de la matriz de covarianzas, sin usar la función `cor` de ?


Indicaciones.

- (a) Para leer los datos almacenados en un fichero de texto se emplea la función `read.table`, como se indica en la página 9 de los apuntes. Se supone que los datos se encuentran dispuestos en el fichero, de manera que las filas son los individuos y las columnas son las variables. Con la función `read.table` se crea un `data.frame` con la misma estructura de individuos por variables. La opción `header=TRUE` hace que la primera línea del fichero texto se interprete como nombres de las variables.
- (b) Este apartado se resuelve utilizando la función `colMeans`, aplicada al `data.frame` obtenido en el apartado anterior.
- (c) La matriz de covarianzas se obtiene con la función `cov`, y las submatrices se consiguen indicando los índices de las columnas del `data.frame` que interese extraer, según se indica en la página 9 de los apuntes. La matriz de correlaciones se obtiene con la función `cor`.
- (d) La matriz de correlaciones se puede obtener de la matriz de covarianzas, multiplicando por matrices diagonales con las inversas de las desviaciones típicas de las variables en la diagonal, a ambos lados de la matriz de covarianzas. En esta operación se utiliza la función `diag`.

Actividad 2: En relación con el Ejemplo 2 de los apuntes, crearemos un fichero con los datos (por ejemplo, denominado “hipertension.txt”), y después elaboraremos un *script* de  que permitira realizar las tareas descritas a continuación.

- (a) Leer los datos del fichero y calcular el vector de medias y la matriz de covarianzas muestrales, para el vector formado por la edad y la presión arterial máxima.

- (b) Representar el diagrama de dispersión de las dos variables: edad y presión arterial máxima.
- (c) Obtener la Figura 4 de los apuntes (página 20), esto es, representar en una sola figura cuatro gráficos que representen los diagramas de dispersión de la edad y la presión arterial máxima, con los datos originales, con los datos centrados, con los datos estandarizados de manera univariante y con los datos estandarizados de manera multivariante.

Indicaciones. Para crear el fichero con los datos, podemos usar cualquier editor de texto, incluso el propio editor que proporciona  para los *scripts*. Situiremos los valores de las dos variables en columnas y escribiremos los nombres de las variables en la primera fila. Como orientación, podemos seguir el formato del fichero “familias.txt”. Una vez tecleados los datos, guardaremos el fichero con extensión *.txt*.

- (a) En este apartado se piden los mismos resultados de la actividad anterior, por lo que seguiremos las pautas que se indicaron allí.
- (b) Se resuelve utilizando la función `plot`.
- (c) Para obtener cuatro gráficos en una misma figura (con forma de tabla 2×2) se emplea el comando `par(mfrow=c(2,2))`. Tras ejecutar este comando, los gráficos que se vayan creando se van añadiendo en esa disposición de tabla.

Para obtener estos cuatro diagramas de dispersión se emplea la función `plot`. Como la escala es importante para apreciar el efecto de la estandarización univariante, en estos cuatro gráficos emplearemos la opción `asp=1` dentro de la función `plot`, lo cual hace que se utilicen las mismas escalas en los ejes horizontal y vertical. El punto que representa el vector de medias se obtiene con la función `points`.

Los datos centrados se consiguen multiplicando el `data.frame` a la izquierda por la matriz siguiente (donde n es el tamaño muestral):

$$M = \begin{pmatrix} 1 - 1/n & -1/n & \cdots & -1/n \\ -1/n & 1 - 1/n & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1/n \\ -1/n & \cdots & -1/n & 1 - 1/n \end{pmatrix}$$

que se conoce como matriz centradora. Para que el `data.frame` pueda participar en un producto de matrices, lo convertimos en matriz con la función `as.matrix`. La matriz centradora se puede crear con la ayuda de la función `diag` para los unos de la diagonal, tras lo cual bastaría con restar el escalar $1/n$, que de ese modo se restaría automáticamente en todos los elementos de la matriz.

Los datos estandarizados de manera univariante se obtienen multiplicando los datos centrados por la derecha por una matriz diagonal con las inversas de las desviaciones típicas de las dos variables.

Los datos estandarizados de manera multivariante se obtienen multiplicando los datos centrados por la derecha por la matriz $S^{-1/2}$, siendo S la matriz de covarianzas. Para calcular la matriz $S^{-1/2}$ se emplea la representación en autovalores y autovectores de S , como se indica en el apéndice de este tema.