Boletín Tema 1

Tratamiento de Datos. Grado en Ciencia de Datos- UV

Marcelino Martínez

2023-02-03

1. Considera los conjuntos de datos **mammals** del paquete **MASS** y **Animals2** del paquete **robustbase**.

    a. Mira la las características de ambos conjuntos usando la ayuda.

    b. Usa las funciones **dim, head, tail, str** para una primera visión de los conjuntos de datos.

    c. Muestra los nombres de las filas y las columnas (**rownames**, **colnames**)

    d. Usa la función **intersect** y almacena en la variable *commonAnimals* los aminales que aparezcan en ambos conjuntos

    e. Usa **setdiff** para averiguar qué animales no están en ambos conjuntos. ¿Cuántos son ?. ¿Qué tipo de animales son?

    f. Determina las diferencia entre los animales que no aparecen en ambos conjuntos.

2. La funcion **qqPlot** del paquete **car** puede ser utilizada para determinar gráficamente si una serie de puntos siguen una distribución de datos Gaussiana. Si las muestras están dentro de las líneas discontinuas podemos indicar que siguen una distribución Gaussiana con un 95 % de confianza. Utilizando esta función representa el logaritmo neperiano (**log**) del peso del cerebro (**brain weigths**) del registro de datos **mammals** del paquete **MASS** y conjunto de datos **Animals2** de la librería **robustbase**. ¿Presentan el mismo comportamiento ?.¿Podríamos decir que siguen una distribución Gaussiana ?

3. La función **library** sin argumentos abre una ventana y muestra las librerías que han sido instaladas.

    a. Asigna el valor devuelto por esta función a la variable **libReturn** y observa su estructura.
    b. Uno de los elementos de la lista es una matriz de caracteres. Muestra por pantalla los 5 primeros elementos de esta matriz usando la función **head**.
    c. Determina el número de librerías que tienes instaladas.

4. En las transparencias del tema 1 se citan los primeros pasos a seguir cuando se analiza un nuevo conjunto de datos.

    a. Determina las tres primeras etapas para el conjunto de datos **cabbages** del paquete **MASS**
    b. Puedes determinar el número de valores perdidos (almacenados como **NA** en R) usando la función **is.na**. Determina el número de valores perdidos para cada una de las variables del conjunto **cabbages**.
    c. Repite los apartados anteriores con el conjunto de datos **Chile** del paquete **carData**.
    d. Utiliza la función **summary**, sobre **cabbages** y **Chile** y observa como, además de otros estadísticos, también devuelve el número de valores perdidos de cada variable.

5. Muchas pruebas estadísticas suponen que los datos siguen una distribución Gaussiana. Utiliza la aproximación visual proporcionada por **qqPlot** para determinar si podemos asumir que las variables **HeadWt** y **VitC** del conjunto **cabbages** verifican esta condición.

6. Una representación habitual, para determinar la distribución de los datos de una variable cuantitativa es el histograma (**hist**). Determina, de forma aproximada, utilizando el histograma, si hay diferencias entre los contenidos de vitamina C (**VitC**), para las diferentes variedades de calabaza (variable **Cult**), en el conjunto de datos **cabbages**.

7. Un modelo sencillo para relacionar variables es la *predicción lineal*. En el siguiente ejemplo se utiliza el conjunto de datos **whiteside**, de la librería **MASS**. Esta aproximación propone un modelo que predice una variable a partir de otra. Una primera etapa para plantear esta aproximación sería representar ambas variables mediante un diagrama de dispersión (Gráfico XY) y determinar si la relación entre variables "parece" lineal. Si es así, podemos plantear un modelo lineal (en este caso según un factor), donde se aprecia claramente que existe una relación lineal entre las dos variables consideradas. Observa y ejecuta el siguiente código.

```
#Diagrama de dispersión global.
plot(whiteside$Temp, whiteside$Gas)
#Diagrama de dispersión etiquetando según un factor.
plot(whiteside$Temp, whiteside$Gas, pch=c(6,16)[whiteside$Insul])
legend(x="topright",legend=c("Insul = Before","Insul = After"), pch=c(6,16))
# Planteamos 2 modelos lineales, uno para los datos de cada factor
Model1 <- lm(Gas ~ Temp, data = whiteside, subset = which(Insul == "Before"))
Model2 <- lm(Gas ~ Temp, data = whiteside, subset = which(Insul == "After"))
# Representamos las rectas correspondientes a cada modelo lineal
abline(Model1, lty=2)
abline(Model2)
```

a. Utiliza un procedimiento análogo para determinar si se aprecia una relación lineal entre los niveles de vitamina C, **VitC** en función del peso de la calabaza, **HeadWt**, en el conjunto de datos **cabbages**.

b. Repite el apartado anterior, pero obteniendo un modelo para cada una de las dos variedades de calabaza, **Cult**. VerParámetros básicos plot.

c. Usa **summary** con cada uno de los modelos obtenidos y observa **Coefficients**. Dado que hemos planteado un modelo $y = mx+n$, donde $y = VitC$ y $x = HeadWt$. La función **lm** nos permite obtener **(Intercept)**; **n** y la pendiente **HeadWt**; **m** (además de otros parámetros adicionales que evalúan la caracterísiticas del modelo). Observa que en todos los casos, la pendiene es negativa indicando que las calabazas de más peso contienen menos vitamina C. No te preocupes por el resto de parámetros del modelo, por el momento.