Problem Set II

Nicolás Forteza

Noviembre 22'

E-mail: nicolas.forteza@bde.es
Lugar: *online*

Tutorías: Después de clase, o bajo cita
Class Hours: L-X-J, 19:00-21:00

1-. Matrices y primeros pasos en un análisis de datos.

Vamos a crear matrices de $2x3$, pero con diferentes datos:

1-. Crea una matriz vacía.

2-. Crea una matriz de números aleatorios. Guárdala en una variable.

3-. Accede a la matriz anteriormente guardada. En concreto, a lo siguiente:

- Celda a_{11}
- Celda a_{22}
- Segunda columna.
- Tercera fila.
- Diagonal de la matriz.

4-. Crea otras matrices de números aleatorios, pero esta vez con las dimensiones necesarias o sufientes como para poder realizar las siguientes operaciones:

- Suma
- Resta
- División
- Multiplicación
- Elevar a un exponente.

Calcula además las dimensiones de cada matriz resultante.

5-. Comprueba si la matriz creada en el ejercicio 2 es invertible o no. Calcula la inversa si lo fuera.

6-. ¿Es la matriz del ejercicio anterior una matriz ortogonal?

7-. La regresión lineal es una de las técnicas estadísticas más utilizadas para entender y modelizar la relación entre dos o más variables. Hay dos tipos de variables en concreto que usamos en un análisis de regresión lineal:

- Variable dependiente/variable a explicar/variable objetivo: como su nombre indica, es la variable que nos interesa explicar en función de otra/s variable/s.

- Variable/s independiente/s o explicativa/s: son aquellas variables que explican la variable dependiente; mediante la observación de éstas variables, podemos inferir, explicar, predecir la variable dependiente.

Cuando la variable objetivo depende linealmente de la variable independiente, se dice que podríamos modelar la variable objetivo con una regresión lineal. O en otras palabras, la variable objetivo es una función lineal de la variable independiente. En concreto, el modelo suele tener la siguiente forma:

$$Y = \alpha + \beta X + \epsilon$$

Y es la variable a explicar, α es una constante, β son los coeficientes del modelo que junto con X (mi matriz de variable o variables explicativas) y ϵ (un error que los datos son incapaces de explicar), forman nuestra explicación o predicción.

No vamos a entrar en cómo se resuelve dicha ecuación (hay que hallar α, β y ϵ). La fórmula para hallar los coeficientes es la siguiente:

$$\beta = (X^T X)^{-1} X^T Y$$

Como véis, es un conjunto de multiplicaciones de matrices.

Vamos a realizar una regresión lineal con un conjunto de datos que R tiene en su entorno base.

- Ejecutar primero el comando `data("USArrests")`. Esto carga un dataframe, que ya vamos a ver qué es en lecciones futuras. A continuación, ejecutar `force("USArrests")`.
- Con el comando `View`, inspeccionar el objeto `USArrests`.
- Ver cuál es la dimensión de `USArrests`.
- Guardar la primera columna en la variable `y`.
- Guardar la segunda columna en la variable `x`. ¿Qué significa esto?.
- Vamos a realizar un gráfico por primera vez. Ejecutar el comando `plot(x, y)`. ¿Se ve alguna relación?.
- Antes de hacer el modelo de regresión lineal, tenemos que añadir una columna a nuestro vector-columna X. Esta columna tiene que ser de 1 (será nuestro α de la ecuación). Guardar el resultado en `x`.
- Realizar el modelo de regresión lineal. ¿Cuáles son los coeficientes? Guárdalos en otra variable llamada `coef`.
- Calcular el error.
- Calcular la recta de regresión y el error. Graficar la recta de regresión, e interpretar el modelo.

2-. Definiciones, bucles, etc.

1-. Crea funciones en R para calcular lo visto en el ejercicio anterior. Es decir, crea funciones para calcular los coeficientes de un modelo de regresión lineal y el error.

2-. Crea una función que realiza un gráfico de puntos o *scatter plot*, y que recibe como argumentos dos vectores. Especificar en estos argumentos, que si no se le pasa nada, por defecto grafique la función dos vectores de números aleatorios.

3-. Usa un bucle para crear una función llamada `num.odd`, que devuelve cuantos números impares hay en un vector `v`. Después, escribe una función igual, pero sin usar bucles.

4-. Los dos primeros términos de una secuencia de Fibonacci son 1. Los términos siguientes de la secuencia se encuentran sumando los dos términos anteriores. Escribe una función `fibonnaci(n)` que produce los primeros n términos de Fibonnaci.

5-. Escribe una función `run.length(p)` que simula la tirada de una moneda (que puede estar sesgada) con probabilidad p, de caer cara o cruz, y devuelve el número de cruces que han salido hasta que ha salido cara. Después, lanza la funcion 100 veces y encuentra el máximo de cruces que han salido entre las 100 simulaciones.

6-. El Teorema Central del Límite es un teorema central en estadística que garantiza que la distribución de las medias muestrales de una distribución estadística siempre sigue una distribución normal. Crea un vector de datos de una distribución no normal (por ejemplo, de una distribución uniforme con `runif()`), de longitud igual a 1000, y después computa la media de 100 elementos aleatorios de ese vector uniforme 100 veces. Dibuja la distribución de estas 100 medias con `hist`.