7.4.1.2 Media aritmética, varianza, desviación, covarianza, media móvil…

Los cálculos estadísticos más básicos incluyen estimadores de la media, la deviación, etc., partiendo de los datos de una muestra de una población dada.

Su significado y las fórmulas que conducen a su cálculo son muy sencillas y se habrán estudiado previamente en la asignatura de estadística (así que solo repetiremos estas últimas como recordatorio).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

siendo:

x : representa a la variable aleatoria.
n : el número de datos de la muestra.
$\bar{x}$: el valor de la media de x.
s_x^2: el valor de la varianza muestral de x.
s_x : el valor de la deviación típica o estándar muestral de x.

Normalmente, en tareas de computación, la desviación estándar no se calcula con la fórmula anterior sino con la siguiente (es una expresión que permite realizar el cálculo sin tener que almacenar los valores de x).

$$s_x = \sqrt{\frac{(\sum_{i=1}^{n} {x_i}^2) - n * \bar{x}^2}{n-1}}$$

Otro cálculo habitual es la covarianza muestral entre *n* valores de dos variables aleatorias *x* e *y*.

$$s_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{n-1}$$

Y su fórmula computacional de cálculo (sin tener que almacenar las variables *x* e *y* en arrays dinámicos).

$$s_{x,y} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\,\bar{x}\,\bar{y}}{n-1}$$

Cuando nos encontramos con series temporales (conjuntos de datos donde el valor de la variable aleatoria depende también del tiempo), se usan habitualmente otras dos definiciones de media.

La media móvil simple consiste en promediar los últimos *k* valores en el tiempo de *x*.

$$\bar{x}_t = \frac{\sum_{i=t-k+1}^{t} x_i}{k}$$

Es un valor que evoluciona a lo largo del tiempo, descartando ("olvidando") progresivamente los valores más antiguos y centrándose en los *k* más nuevos.

Pero el cálculo de esta fórmula requiere almacenar las *x* en una estructura de tipo cola FIFO para evitar malgastar mucha memoria. Para sortear este problema, hay muchos algoritmos que prefieren implementar otra variedad de media que se llama media móvil acumulada.

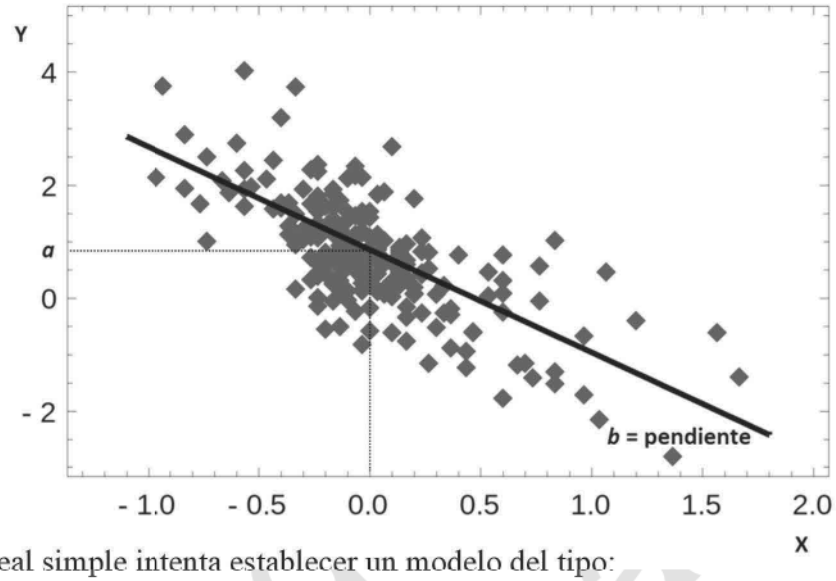
$$\bar{x}_{t+1} = \alpha\, x_{t+1} + (1-\alpha)\,\bar{x}_t$$

donde la letra *t* representa el instante en el que se ha obtenido el dato de la muestra. $\bar{x}_{t+1}$ es la media móvil acumulada en el instante *t* + 1 y $\bar{x}_t$ la media móvil acumulada calculada en el instante *t*. α es la "tasa de innovación"; cuanto más grande sea, más pondera el último valor de *x* y menos los valores antiguos de *x*. Así pues, es posible establecer un compromiso sobre qué valores a lo largo del tiempo consideramos más importantes para evaluar la media. En este último algoritmo tampoco es necesario almacenar los valores de *x*.

7.4.1.3 Regresión lineal

Para comprender este apartado se supone que el alumno ha cursado previamente una asignatura básica de estadística.
Solo se va a abordar la programación en C de una regresión lineal simple con dos variables aleatorias:

- Una a la que llamaremos dependiente (y).
- Y otra que será la independiente (x).



Se conoce que una regresión lineal simple intenta establecer un modelo del tipo:

$$y = a + b\,x + \varepsilon$$

donde a (ordenada en el origen) y b (pendiente de la recta) son los coeficientes a calcular de la regresión lineal, y ε es un error aleatorio.

El cálculo de a y b se puede realizar mediante las fórmulas siguientes:

$$b = \frac{s_{x,y}}{s_x^2}$$

$$a = \bar{y} - b\,\bar{x}$$

En el caso de una regresión lineal múltiple, la ecuación es similar a la de la regresión lineal simple:

$$Y = A + X\,B + \epsilon$$

Pero ahora Y es un vector columna con los valores la variable dependiente para cada caso, A es otro vector columna con el parámetro de la ordenada en el origen a calcular (a) replicado en tantas filas como casos existan, X es una matriz de tantas filas como casos existan y tantas columnas como variables independientes existan y, por último, B es otro vector columna de parámetros a calcular con tantas filas como variables dependientes existan.

Puesto de otra forma (siendo p el número de variables independientes):

$$y_j = a + \sum_{i=1}^{p} x_{j,p}\,b_p + \varepsilon_j\ , \quad 1 \le j \le n$$

La solución, mediante el método de minimización del error medio cuadrático, para este sistema de ecuaciones da:

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

$$a = \bar{y} - \bar{X}\,B$$

donde $\bar{X}$ es un vector de tipo fila en el que cada elemento es la media de los valores de cada columna de X.