Tema 2. Estadística descriptiva de dos variables

1.Conceptos básicos

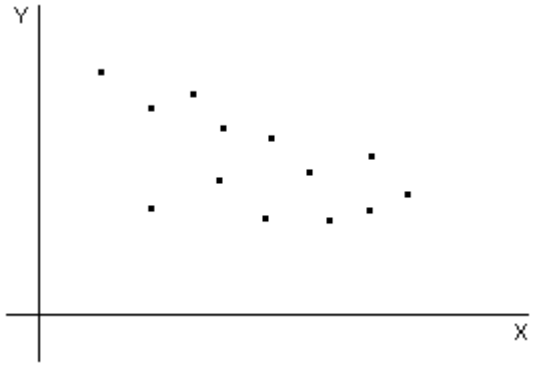
1.1 Planteamiento

Para hacer el estudio de dos variables cuantitativas X e Y, supondremos que disponemos de una muestra de n pares de observaciones de X e Y

$$(x_1, y_1),...,(x_n, y_n)$$

Los cálculos de las medias y varianzas para cada variable aleatoria son idénticos a los ya estudiados en el tema 1, por lo que es fácil obtener $\bar{x}$, v_x, $\bar{y}$, v_y.

La representación de los datos se obtienen del clásico diagrama XY obteniendo una nube de puntos que nos da una idea visual de las posibles relaciones existentes entre las dos variables.



1.2 Covarianza muestral

Se representa por $\text{cov}_{x,y}$ y se define como

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

aunque es usual utilizar la fórmula alternativa

$$\text{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

La covarianza nos dará idea de la relación que existe entre las dos variables y que está relacionada con la forma de la nube de puntos.

2. Modelo de regresión lineal

Una vez obtenida la nube de puntos es interesante encontrar la recta que mejor la represente. A esta recta la llamaremos **recta de regresión** y matemáticamente desearemos que dicha recta minimice la media de cuadrados de las distancias de los puntos a la recta.

Esta distancia puede medirse horizontalmente o verticalmente. En esta asignatura nos interesará minimizar las distancias en sentido vertical, y denominaremos a dicha recta, recta de regresión de Y sobre X.

2.1 Recta de regresión de *Y* sobre *X*

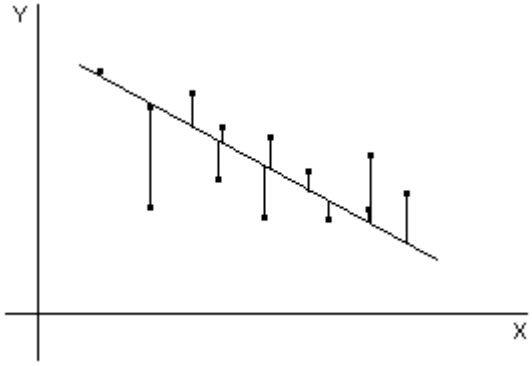
La recta de regresión de *Y* sobre *X* es la recta de la forma $y = a + bx$, que minimiza el error cuadrático medio (E.C.M.):

$$E.C.M. = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

y es

$$y - \bar{y} = \frac{\text{cov}_{x,y}}{v_x}(x - \bar{x})$$

Gráficamente es



2.2 Varianza residual

La **varianza residual** es el error cuadrático medio que se comete en la recta de regresión de *Y* sobre *X* y se define como

$$\text{Varianza residual} = v_y\left(1 - \frac{(\text{cov}_{x,y})^2}{v_x v_y}\right)$$

El cociente que aparece en la fórmula de la varianza residual recibe un nombre específico tal y como veremos en el siguiente apartado.

2.3 Coeficiente de correlación muestral

El **coeficiente de correlación muestral** entre X e Y se representa por *r* y se define como

$$r = \frac{\text{cov}_{x,y}}{\sqrt{v_x v_y}}$$

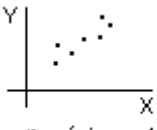
Esto permite reescribir la fórmula de la varianza residual utilizando el coeficiente de correlación muestral. Así

$$\text{Varianza residual} = v_y\left(1 - r^2\right)$$

El coeficiente de correlación muestral siempre tiene un valor entre –1 y 1. Su signo dependerá exclusivamente de la varianza (ya que v_x y v_y son siempre positivos). Además, un valor cercano a 1 (ó a –1) nos indicará que la nube de puntos es próxima a una recta, mientras que si es próximo a 0 nos indica una dispersión general de la nube de puntos.

r próximo a -1

r próximo a 0



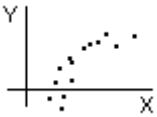
r próximo a 1

3.Aplicaciones del modelo de regresión lineal

No siempre la nube de puntos se aproxima a una recta. En algunas ocasiones se asemeja más a otras funciones como las exponenciales o logarítmicas.

3.1 Regresión logarítmica

Cuando la nube de puntos es como la de la siguiente figura



se aconseja ajustar mediante un modelo de la forma

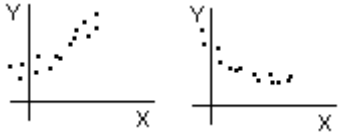
$$y = a + b \log x$$

(donde $\log x$ es la función logaritmo neperiano)

En este caso, hacemos un cambio de variable $T = \log x$ y hallamos la recta de regresión de Y sobre T. Sólo nos queda deshacer el cambio, es decir, de nuevo $T = \log x$.

3.2 Regresión exponencial

Cuando la nube de puntos es como las de la siguiente figura



se aconseja ajustar mediante un modelo de la forma

$$y = ae^{bx}$$

En este caso, se toman logaritmos

$$\log y = \log a + bx$$

Si ahora llamamos $T = \log y$, sólo tenemos que hallar la recta de regresión de T sobre X y deshacer el cambio.