NÚMEROS. Revista de didáctica de las matemáticas
Volumen 35, septiembre de 1998, páginas 32-43

Aspectos geométricos de la regresión y correlación lineal

Juan-Bosco Romero Márquez,
María de los Ángeles López y Sánchez-Moreno

Resumen

Con este trabajo elemental de investigación en el aula presentamos como una experiencia el estudio del problema: LA REGRESIÓN Y LA CORRELACIÓN LINEAL DE UNA VARIABLE ESTADÍSTICA BIDIMENSIONAL Y EL SIGNIFICADO GEMOÉTRICO, utilizando para ello las propiedades elementales del trinomio de segundo grado y su gráfica, que es una parábola, y la traslación de ejes de coordenadas, evitando con ello el cálculo con derivadas.

Al final damos una aplicación a la geometría del plano.

Abstract

In this article we presents some geometric topics over the regression and correlation lineal in the secondary teaching.

Introducción

Comenzamos enunciando el siguiente problema.

Problema.- Sea la variable estadística bidimensional (X,Y) donde X e Y son variables estadísticas unidimensionales, de la que conocemos la distribución conjunta de puntos, $P_i(x_i,y_i)$, con i = 1,... n

Queremos encontrar e interpretar las rectas de regresión lineal de y sobre x, dada por y = b x + a, y la recta de regresión lineal de x sobre y, dada por x = b' y + a', utilizando como criterio de optimización el de mínimos cuadrados, para determinarlas como las curvas de ajuste o aproximación de la nube de puntos P_i, i = 1,... n.

Es decir: calcular las rectas de regresión lineal de tal forma que la desviación cuadrática, de y sobre x, como de x sobre y, que están dadas, respectivamente, por las fórmulas:

$$S(a, b) = \sum_{i=1}^{n} (y_i - (a+b\, x_i))^2 \quad \textit{(1)}, \text{ (y sobre x), } \quad \text{y}$$

$$S(a',b') = \sum_{i=1}^{n} (x_i - (a'+b'\,y_i))^2 \quad \textit{(2)}, \text{ (x sobre y)},$$

sean mínimas.

Basta resolver el problema para el caso *(1)*, ya que el caso *(2)*, se resolverá por simetría intercambiando los papeles de las variables.

Observación: A la desviación cuadrática definida anteriormente, también se le llama error cuadrático.

Resultados previos

Los alumnos deben conocer y manejar completamente los siguientes conceptos y resultados:

a) La función cuadrática $y = f(x) = ax^2 + bx + c$, donde $a,b,c, \in \mathbb{R}$, con a no nulo y cuya gráfica es una parábola con eje de simetría vertical, y con las ramas hacia arriba (cóncava) si $a > 0$, y con las ramas hacia abajo (convexa), si $a < 0$.

Los resultados anteriores son consecuencia del estudio del signo de la función cuadrática, $y = ax^2 + bx + c$, y de sus propiedades algebraicas y geométricas elementales, en su dominio, $\mathbb{R}$

a_1) El vértice, V, de la parábola viene dado por:

V (-b/2a, f(-b/2a)), que es, el punto mínimo o máximo absoluto de la función o de la parábola, respectivamente, para $a > 0$, y $a < 0$.

a_2) Dada la variable estadística X, cuyos valores dados son x_i, $i = 1,2,...n$.

Definimos $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, *(3)*, la media aritmética de los datos, x_i, $i = 1,2,...n$.

Definimos el error de un dato x_i, con respecto a la media $\overline{x}$, o con respecto a un dato arbitrario x, por las expresiones:

$d_i = x_i - \overline{x}$, o, $d_i\,(x)\ 0\ x_i = x$. Es claro que puede ocurrir que la suma de todos los errores anteriores puede dar cero. Por esta causa esta suma de errores o desviaciones no es significativa para estudiar las propiedades de la media aritmética, $\overline{x}$.

Para caracterizar la media aritmética, $\overline{x}$, necesitamos definir el error cuadrárico de los errores $d = x - \overline{x}$, $i = 1,2,...n$, de las desviaciones de los datos x_i, con respecto a la media, $\overline{x}$, como sigue: