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:

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

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

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 $\bar{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, *(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 $\bar{X}$, o con respecto a un dato arbitrario x, por las expresiones:

$d_i = x_i - \bar{X}$, 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, $\bar{X}$.

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

$$S(x) = \sum_{i=1}^{n} (x_i - \bar{x})^2, \text{ donde } x \in \mathbb{R}, \textit{(4)}$$

Es claro, que la función $y = S(x)$ es una función cuadrática de la variable x, que es, positiva, y que, por lo tanto alcanza su mínimo absoluto en el punto $\bar{x} = \frac{\sum x_i}{n}$, es decir en la media.

a) Traslación de ejes de coordenadas.- Si designamos por XY, y X'Y' dos sistemas de coordenadas cartesianas relacionados por el vector de traslación, $t = (a,b)$ tenemos que, si $P(x,y)$ y $P(x',y')$, son las coordenadas del punto P, con respecto a los ejes XY, y X'Y', respectivamente, tenemos para las ecuaciones de la traslación de ejes:

$$x' = x - a, \quad y' = y - b. \textit{(5)}$$

Con todo lo anterior ya estamos en condiciones de resolver el problema propuesto.

Rectas de regresión lineal. Coeficiente de correlación lineal y su interpretación geométrica y algebraica.

Sea la variable estadística bidimensional, (X,Y), de la que conocemos la distribución de datos, $P_i(x_i,y_i)$, $i = 1,2,... n$.

Llamamos centro de gravedad o baricentro de la distribución o nube de puntos anterior, al punto $G(\bar{x}, \bar{y})$, donde $\bar{x}$ e $\bar{y}$, son las medias aritméicas de las variables estadísticas unidimensionales x e y, respectivamente.

Recuérdese que, en cierto sentido, toda variable estadística bidimensional se puede considerar como la proyección o recomposición según nos convenga, de las variables estadísticas, X e Y, respectivamente.

Vamos a determinar, por ejemplo, la recta de regresión de Y, sobre X, es decir, la recta de la forma: $y = a + b x$, (donde a y b son los parámetros a determinar del haz de rectas del tipo anterior), de manera que el error cuadrático, de y sobre x, dado por (1), sea mínimo. Esto es: la recta de y sobre x, que está más cerca de la nube de puntos de acuerdo con el criterio del error cuadrático, o el también llamado "método de ajuste o de aproximación por nínimos cuadrados".

Para resolver el problema propuesto definimos la traslación de ejes, $x' = x - \bar{x}$, $y' = y - \bar{y}$, de modo que, con este cambio de coordenadas, el origen del nuevo sistema de coordenadas, es el punto 0' $(\bar{x}, \bar{y})$.

Así, en el nuevo sistema de coordenadas la variable bidimensional (X,Y) se transforma en la variable bidimensional (X',Y'), de tal manera que:

$$x'_i = x_i - \bar{x}, \qquad y'_i = y_i - \bar{y}, \qquad i = 1,2,...n.$$

son las coordenadas de los puntos, P_i, en el sistema de coordenadas, X'Y'.

Por tanto, de acuerdo con la fórmula del error de minimos cuadrados dada por (1), podemos escribir operando y reagrupando términos en la misma que:

$$S(a, b) = \sum (y_i - (a + b\, x_i))2 = \sum (y_i + \bar{y} - \bar{y} - a - b(x_i - \bar{x} + \bar{x})) =$$

$S(a, b) = \sum (y'_i + \bar{y} - a - b\, x'_i - b\, \bar{x})^2$ ***(6)*** , a, b $\in \mathbb{R}$ que es una función de las variables reales, a y b, y que representaría, en el sistema de coordenadas, (a,b,S), un paraboloide elíptico, como es fácil de probar.

Ahora bien, si elegimos entre las rectas que buscamos aquellas en las que $\bar{y} = a + b\, \bar{x}$, es decir las que pasan por el punto $G(\bar{x}, \bar{y})$, que es el centro de gravedad de la distribución, en la relación ***(6)***, obtenemos:

$$S(a, b) = S'(b) = \sum (y'_i - b\, x'_i)^2 = b^2 \sum x'^2_i - 2\, b \sum x'_i\, y'_i + \sum y'^2_i =$$

$= A\, b^2 - 2\, B\, b + C$, ***(7)***, que es una función cuadrática en b, que puede ser representada como una parábola en el plano (b S'), y donde hemos puesto como $A = \sum x'^2_i$, $B = \sum x'_i\, y'_i$ y $C = \sum y'^2_i$

Como $A > 0$, el vértice de ***(7)*** está en el mínimo absoluto, y viene dado por:

$$\bar{b} = \frac{2\, B}{2\, A} = \frac{\sum x'^2_i\ y'^2_i}{\sum x'^2_i} \qquad \textbf{(8)}$$

De la relación ***(8)***, obtenemos que el mínimo de ***(7)***, es:

$$S'(b) = \frac{\sum x'^2_i \cdot \sum y'^2_i - (\sum x'_i\, y'_i)^2}{\sum x'^2_i} \geq 0, \qquad \textbf{(9)}$$

De la fórmula ***(9)***, obtenemos a partir de su numerador positivo una nueva demostración de la desigualdad de Cauchy-Schawarz, y con la igualdad alcanzada, si y sólo si, $y'_i = k\,x'_i$, $i = 1,2,...n$, es decir si los vectores $x' = (x'_1,...,x'_n)$ e $y' = (y'_1,...y'_n)$ son proporcionales.

Por todo lo anterior, tenemos que la recta de regresión de y sobre x viene dada por las dos relaciones siguientes:

$$y = a + b\,\bar{x}, \qquad \bar{y} = a + b\,\bar{x}.$$

Restando ambas, llegamos a

$$y - \bar{y} = b_{y\,x}\,(x - \bar{x}) = \frac{\sum x'_i\ y'_i}{\sum x'^{\,2}_i}\ (x - \bar{x}),$$

donde, $b_{y\,x} = \dfrac{s_{xy}}{s_x^{\,2}}$, se llama coeficiente de regresión lineal de y sobre x,

s_{xy} = C(X,Y) = covarianza de X e Y = $\sum x'_i\,y'_i$, y $s_x^{\,2} = \sum x'^{\,2}_i$, es el cuadrado de la covarianza, de la variable X.

Por simetría se obtiene la recta de regresión de x sobre y. Viene dada por:

$$x - \bar{x} = b_{xy}\,(y - \bar{y}),$$

donde b_{xy}, es el coeficiente de regresión lineal de x sobre y, y viene dado por

$b_{xy} = \dfrac{s_{xy}}{s_y^{\,2}}$, con las mismas notaciones y definiciones anteriores.

Se define el coeficiente de correlación lineal, $r = \pm\sqrt{b_{yx}\ \ b_{xy}}$, media geométrica con signo de b_{yx} y b_{xy}.

Se puede demostrar que:

1) $-1 \le r \le 1$.

2) $r^2 = \dfrac{\text{variación explicada}}{\text{variación total}} = \dfrac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2} = \dfrac{s_{xy}}{s_x^{\,2}\ s_y^{\,2}}$

3) Definimos como $s_{y \cdot x}$ = error .típico de estima de y sobre

$$x = \sqrt{\frac{\sum (y - y_{est})^2}{n}} \quad , \quad y$$

$s_{x \cdot y}$ = error típico de estima x sobre y $= \sqrt{\frac{\sum (x - x_{est})^2}{n}}$

Demostrar que

i) $s_{y \cdot x}^2 = s_y^2 (1 - r^2) = S' (\bar{b})$ y $s_{x \cdot y}^2 = s_x^2 (1 - r^2)$.

ii) $s_{y \cdot x}^2 / s_{x \cdot y}^2$ para la regresión lineal. ¿Es el resultado cierto para la regresión no lineal?

Observación y Problemas.

1) La resolución clásica del problema de buscar la recta de regresión lineal de y sobre x, dada por $y = a + b\,x$, minimizando la función

$S(a,b) = \sum_{i=1}^{n} (y_i - (a + b\,x))^2$, como una función diferenciable en las variables , a y b, utilizando el cálculo diferencial de dos variables para la determinación de su mínimo, ver (1).

2) Una buena colección de ejercicios de ampliación para profesores y alumnos sobre los siguientes tópicos de este tema: rectas de regresión lineal, coeficientes de regresión, coeficiente de correlación lineal, coeficiente de determinación (r^2), varianza residual, con el estudio y el significado de sus propiedades algebraicas y geométricas, tanto en lo que se refiere a la población, como a la muestra se puede enontrar en los libros citados en la bibliografía, y en especial, en (1).

También se encuentran en todos los libros citados en la bibliografía la representación gráfica e interpretación de las rectas de regresión lineal, y de sus topicos asociados.

En (1), también se estudia la regresión no lineal y la regresión múltiple y parcial, lineal y no lineal.

Problemas

1) Si la recta de regresión de mínimos cuadrados de y sobre x viene dada por

$$y = a + b\,x, \text{ demostrar } s^2_{x\cdot y} = \frac{\sum y^2 - a\sum y - b\sum x\,y}{n}$$

2) Utilizando 1), probar:

$$s^2_{y\cdot x} = \frac{\sum (y-\bar{y})^2 - b\sum (x-\bar{x})(y-\bar{y})}{n}$$

3) Demostrar que $\sum (y-\bar{y})^2 = \sum (y-y_{est})^2 + \sum (y_{est}-\bar{y})^2$

4) La recta de regresión de y sobre x es $y = a + b\,x$ donde $b = r\,s_y / s_x$. Análogamente, la recta de regresión de x sobre y es $x = c + d\,y$ donde $d = r\,s_y / s_x$. Entones $b\,d = r^2$.

5) Aplicando la fórmula del ámgulo entre dos rectas, demostrar que el ángulo determinado por las dos rectas de regresión lineal muestrales de mínimos cuadrados viene dada por

$$\text{arct}\ \frac{(1-r^2)\,s_{xy}}{r^2(s_x^2 + s_y^2)}$$

Aplicaciones y complementos

En esta sección damos diferentes aplicaciones y complementos relacionados con la regresión y correlación lineal.

Aplicaciones

Vamos a resolver un problema de regresión y correlación lineal a través de un problema de tipo geométrico, y cuyo enunciado es el siguiente:

Considera el triángulo rectángulo de vértices A(0,0), B(a,0) y C(0,b). Interpretando estos vértices como datos una variable estadística bidimensional Z = (X,Y), demostrar que las rectas de regresión lineal de Y sobre X, y de X sobre Y son, las medianas correspondientes a los catetos del triángulo.

Calculad el coeficiente de correlación lineal.

Veamos la solución. Podemos considerar los casos: 1) $a > 0$ y $b > 0$; 2) $a > 0$ y $b > 0$. Basta con resolver el primero. Ver la figura (1) aparte.

Tenemos para el cálculo la siguiente tabla de frecuencias.

X	Y	X^2	Y^2	XY
0	0	0	0	0
a	0	a^2	0	0
0	b	0	b^2	0
a	b	a^2	b^2	0

Los momentos de primer y segundo orden de la distribución (X,Y) vienen dados por:

$a_{10} = \bar{x} = a/3$ y $a_{01} = \bar{y} = b/3$, $a_{20} = a^2/3$ y $a_{02} = b^2/3$, $a_{11} = 0$

El centro de gravedad o de la distribución es el baricentro del triángulo, $G(\bar{x},\bar{y})$.

Las desviaciones típicas de X, de Y, y la covarianza de (X,Y) respectivamente, están dadas por:

$$s_x = \sqrt{a_{20} - a_{10}^2} = \frac{a}{3}\sqrt{2}.$$

$$s_y = \sqrt{a_{02} - a_{01}^2} = \frac{b}{3}\sqrt{2}.$$

$$s_{xy} = a_{11} - a_{10}\, a_{01} = -ab/9$$

Nota: Los momentos respecto al origen de una variable estadística bidimensional, Z = (X,Y) de frecuencias unitarias vienen dados por la siguiente expresión:

$$a_{ij} = \sum_{i,j} x^i y^j, \quad \text{donde} \quad i, j = 0, 1, 2 .....$$

Los coeficientes de regresión de Y sobre X, y de X sobre Y, denotados, b_{yx}, b_{xy}, respectivamente son:

$$b_{yx} = s_{xy} / s_x^2 = -b/2a, \quad b_{xy} = s_{xy} / s_y^2 = -a/2b.$$

De aquí, las rectas de regresión de Y sobre X y de X sobre Y, son las siguientes:

$$r_{yx} : y - \bar{y} = b_{yx}(x - \bar{x}) ; \; y - b/3 = -b/2a\,(x - a/3),$$

$$r_{xy} : x - \bar{x} = b_{xy} (y - \bar{y}) ; \; x - a/3 = -a/2b \; (y - b/3)$$

Se puede comprobar sin ninguna dificultad que son las medianas de los catetos del triángulo rectángulo dado.

El coeficiente de correlación lineal

$$r = \pm \sqrt{b_{yx} \; b_{xy}} = \pm \; 1/2$$

Nota: Se puede generalizar este problema para un triángulo cualquiera, y para algunos tipos de polígonos.

Complementos

En la revista Parábola Vol. 14, n.3, editada por la University of New South Wales, 1978, en las p.p. 33-34, encontramos el problema propuesto n. 377:

Sean x_i, y_i (i = 1,2,...n) números reales tales que

$$x_1 \geq x_2 \geq .... \geq x_n \quad y \quad y_1 \geq y_2 \geq .... \geq y_n$$

Probar que si $z_1, z_2, ... , z_n$ es cualquier ordenación de $y_1, y_2, ... , y_n$

entonces $\sum_i (x_i - y_i)^2 \leq \bar{\times}_i (x_i - z_i)^2$

Ver (11) para encontrar la solución.

En la revista The Mathematical Gazette, n. 77 (1993), aparece la siguiente nota: "Comparing Spearman's rank the product-moment correlation coefficients", el siguiente teorema:

Teorema,. Sean (x_1,y_1), (x_2,y_2)... (x_n,y_n) n $\geq$ 2 datos de una variable bidimensioal con $x_1 < x_2 < .... < x_n$ y $y_1 < y_2 < ... < y_n$ tales que $r_s = 1$. Entonces

$$r \geq \frac{1}{n - 1}$$

Los autores son: C. Bradley y N. Lord. Ver (12).

Más recientemente, en la revista Monthly de la MAA, en Vol. 104, n5, May, 1997, pp. 458-459, tenemos el problema 10.384, propuesto por F. Kemp, en 1994, sobre "A Lower Bound on Correlation" y cuyo enunciado es:

"Supongamos $x_1 < x_2 <... < x_n$ y $y_1 < y_2 <... <y_n$. Definimos el coeficiente de correlación r en la forma usual:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

donde $\bar{x}$ e $\bar{y}$ son las medias aritmética de los x e y, respectivamente. Demostrar que

$$r \geq \frac{1}{n - 1}$$

La resolución que se publica es la dada por R.J. Chapman, de la University de Exeter. Exeter, U.K., y la técnica empleada por este es distinta a la dada anteriormente y consiste en utilizar el espacio euclídeo $\mathbb{R}^n$ en el que se construyen un conjunto C, y unos vectores u, v $\in$ $\mathbb{R}^n$, apropiados. Ver (13).

Conclusiones y comentarios

Está claro que en esta demostración para la obtención de las rectas de regresión lineal de y sobre x, de x sobre y, respectivamente, hemos evitado en todo momento el cálculo de derivadas parciales.

Se puede hacer de esta exposición elemental de la experiencia que hemos presentado aquí, en la forma geométrica y algebraica de la regresión y correlación lineal sin grandes dificultades en el segundo ciclo de la ESO, en el bachillerato LOGSE, y en el actual BUP. Esta experiencia se ha realizado en este año con un provecho aceptable para los alumnos.

Todo esto supone un cambio metodológico y didáctico profundo y posiblemente novedoso e importante en la enseñanza de este tema, debido al carácter elemental de la exposición: tanto de los razonamientos como de las técnicas utilizadas. Todo ello es necesario hacerlo así, en este nivel educativo, para conseguir la mejor educación matemática con nuestros alumnos en su doble vertiente: enseñanza y aprendizaje de calidad.

Esperamos las críticas y las sugerencias constructivas que mejoren nuestra

pequeña investigación, presentada en este artículo. Si, además, logramos no aburrir a los alumnos menudo objetivo habremos conseguido, al hacerles más sencillo y más asequible el aprendizaje de las matemáticas en estos niveles educativos.

Hoy, para los tiempos que corren en la enseñanza ya es mucho.

Bibliografía

(1) Spiegel. M.(1970). *Probabilidades y Estadística,* Schaun, MacGraw-Hill, Bogotá.

(2) Mode, E.B. (1982). *Elementos de Probabilidad y Estadística*, Reverte, S.A. Barcelona.

(3) Nieto de Alba, U. (1980). *Introducción a la Estadística*, Vol. I, II, III, Aguilar S.A., Madrid.

(4) A. Haber y R.P. Runyon. (1973). *Estadística General*, Fondo Educativo Interamericano, México.

(5) L.G. Gotkin y L.S. Goldstein. (1079). *Estadística Decriptiva*, I, II, Texto Programado, Limusa, México.

(6) A. Nortes. (1993). *Estadística Teórica y Aplicada*, PPU, Barcelona.

(7) M.R. Spiegel. (1970). *Teoría y Problemas de Estadística*, Schaum, Mac-Graw-Hill, México.

(8) V.E. Gurman. (1974). *Teoría de Probabilidades y estadística matemática*, Mir, Moscu.

(9) J.J. Thomas. (1980). *Introducción al análisis estadístico para economista,* Marcombo, Barcelona.

(10) D. Downing y J. Clark. (1989). *Statistic the easy way*, Barron's, New York.

(11) Parábola. (1978). Vol. 14, Nº 3, p.p. 33-34, Published the University of New South Wales, Kensington, Australia.

(12) C. Bradley y N. Lord. (1993). *Comparing Spearman's rank and the product-moment correlation coefficients*, Vol. 77, p.p. 84-88.

(13) Problema. (1997). 10.384, Monthly, Vol. 104, Nº 5, May, p.p. 458-459.

(14) R. Courant y F. John. (1976). *Introducción al cálculo y al análisis matemático*, Vol. 1, Limusa, México, .

(15) S. Ríos. (1967). *Métodos estadísticos,* De. Castillo, S.A. Madrid.

(16) W. Feller. (1962). *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons, New York.

(17) H. Cramer. (1953). *Métodos Matemáticos de la Estadística*, Aguilar, S.A., Madrid.

(18) L.J. Kazmier. (1978). *Teoría y problemas de Estadística aplicada a la Economía y a la Administración*, Schaum, MacGraw-Hill, Madrid.

Juan Bosco Romero Márquez (Montilla, Córdoba, 1945). Licenciado en Ciencias Matemáticas, por la Universidad Complutense de Madrid. Fue profesor de dicha Universidad, entre 1971 y 1976. Becario del Consejo Superior de Investigaciones Científicas, desde 1972 a 1975. Actualmente es catedrático de Enseñanza Media en Matemáticas en el I.E.S. "Isabel de Castilla" de Avila. Numerosas publicaciones en artículos, comunicaciones, problemas, en revistas nacionales y extranjeras sobre Matemática Elemental y Superior, sobre todo en problemas. Vicepresidente por Castilla-León de la Sociedad "Puig Adam" de Profesores de Matemáticas.

María de los Ángeles López y Sánchez-Moreno (Madrid, 1946). Licenciada en Ciencias Matemáticas por la Universidad Complutense de Madrid. Desde 1976 es profesora de Matemáticas en diferentes Institutos. Es profesora Tutora del Centro Asociado de la UNED de Avila. Tiene publicaciones sobre Didáctica de las Matemáticas. Ha presentado diversas comunicaciones en diferentes Jornadas, Congresos, etc.