Distribuciones bidimensionales
Recta de regresión

Miguel Galo Fernández

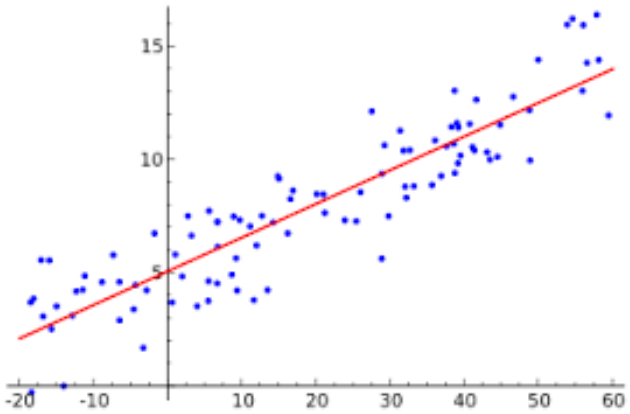


Figura 1: La recta de regresión

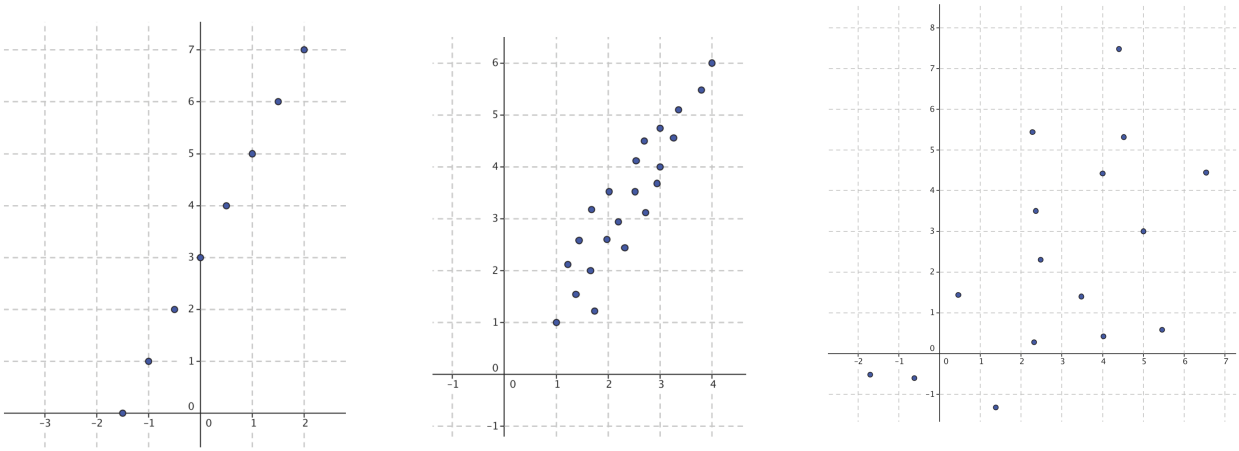
Índice de Contenidos

1. Introducción

Se entiende por conocimiento el acto de conocer, que es la capacidad del hombre para comprender por medio de la razón. Si nos atenemos a esta definición el texto Anaya de Matemáticas orientadas a las Enseñanzas Académicas 4. ESO. (ISBN 9788469810699) no tiene el propósito de transmitir conocimiento en su exposición de la distribución bidimensional. Según se puede observar, no se llega a conseguir transmitir ni una idea intuitiva de lo que se pretende en el tema 10, que es donde se efectúa exposición. ¿Por qué no se dice lo que es la recta de regresión y cuál es su ecuación? ¿Tienen los alumnos conocimientos necesarios para ello? Es evidente que sí, lo demuestro a continuación. Si bien es cierto que un desarrollo en donde se argumente por medio de razonamientos es imprescindible para que los alumnos adquieran conocimiento, no es menos cierto que esas demostraciones no deben ser aprendidas de memoria, basta con que se entiendan. ¿Qué sentido tiene hacer referencia al coeficiente de correlación lineal de Pearson si no se dice lo que es? ¿Por qué se dice que cuanto más próximo a 1 ó a -1 es el coeficiente de correlación lineal mejor se ajusta la nube de puntos a la recta de regresión si no se dice el porqué?, ¿tiene suficientes conocimientos matemáticos el alumno para comprenderlo? Tal y como demuestro a continuación los conocimientos que se precisan para comprender todo esto son muy básicos, lo que no entiendo es como se expone de forma tan lamentable. Y luego nos lamentamos de que hay fracaso escolar, ¡como no va a haberlo!

2. La recta de regresión

Sea una nube de puntos $(x_i, y_i)_{1\leq i\leq n}$ que queremos aproximar a la recta $y = a + bx$. Este es el problema que se plantea con la recta de regresión, cómo hacer esta aproximación. Lo primero que hemos de tener en cuenta es que la recta queda determinada por tan solo dos de los puntos por donde pasa y ya sería casualidad que los puntos de la nube estuvieran alineados, en general la nube de puntos no va a estar compuesta de puntos alineados. La representación gráfica de la nube de puntos presentaría uno de estos tres gráficos:



a b c

En el caso a la nube de puntos está alineada, no hay problema para determinar los coeficientes a y b, basta con elegir dos puntos cualesquiera de la nube y resolver un sistema de dos ecuaciones (una para cada punto que sustituimos en $y = a+bx$) con dos incógnitas (se trata de hallar los valores a y b). **En el caso c** es obvio que no existe recta alguna que pase por todos los puntos de la nube. habría que pensar, en todo caso, en hallar la ecuación de una recta que se aproxime a estos puntos, pero tal y como está la configuración de la nube que nos muestra la gráfica, se hace evidente la imposibilidad de ello. **En el caso b** tampoco hay una recta que pase por todos los puntos pero aquí sí es posible encontrar la ecuación de una recta que se aproxime a la nube de puntos. Para cada punto (x_i, y_i) de la nube hay un punto de la recta $y = a + bx$ al que se aproxima a él. Este punto es $(x_i, a + bx_i)$, llamemos $\hat{y} = a + bx_i$. Al sustituir el punto original de la nube (x_i, y_i) por el punto de la recta al que se aproxima $(x_i, \hat{y}_i)$cometemos un error dado por la expresión $e_i = y_i - \hat{y}_i$. ¿Qué criterio aplicamos para controlar este error? Uno de estos criterios es el de mínimos cuadrados, que consiste en hacer que la suma de estos errores sea mínima, es decir que ha de ser mínima la expresión $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_1)^2 = \sum_{i=1}^{n} y_i^2 - 2y_i\hat{y}_i + \hat{y}_i^2$. Ha de ser mínima la expresión $Q = \sum e_i^2 = \sum y_i^2 - 2\sum y_i(a + bx_i) + \sum(a + bx_i)^2$. Luego:

$Q = \sum y_i^2 - 2a\sum y_i - 2b\sum x_i y_i + \sum(a^2 + 2abx_i + x_1^2 b^2)$ y separando las sumas del último sumatorio queda;

$Q = \sum y_i^2 - 2a\sum y_i - 2b\sum x_i y_i + +na^2 + 2ab\sum x_i + \sum x_i b^2$. Esta expresión la podemos poner como:

$Q = na^2 + 2(b\sum x_i - \sum y_i)a + (b^2\sum x_i^2 - 2b\sum x_i y_i + \sum y_i^2)$, recordemos que el valor de Q ha de ser mínimo. Aplicamos ahora el método ceteris paribus a la anterior ecuación (consiste en variar la incógnita **a**, manteniendo constante el valor de b, o al contrario hacer variar **b** manteniendo constante el valor de **a**).

Si hacemos **b** constante dejando que **a** varíe la expresión de Q es $Q = na^2 + 2la + r$, siendo $l = b\sum x_i - \sum y_i$, $r = b^2\sum x_i^2 - 2b\sum x_i y_i + \sum y_i^2$. Q es una parábola cóncava (ya que $n > 0$) que alcanza un mínimo en su vértice que se alcanza en el punto $a = \dfrac{-2l}{2n} \Rightarrow an + l = 0$. Si sustituimos el valor de l: $an + b\sum x_i - \sum y_i = 0$, dividiendo cada sumando por n y trasponiendo $\sum y_i$ obtenemos la primera ecuación normal $\boxed{a + b\overline{x} = \overline{y}}$

Q lo podemos expresar como $Q = b^2\sum x_i^2 + 2(a\sum x_i - \sum x_i y_i)b + (\sum y_i^2 - 2a\sum y_i + a^2 n)$. Si hacemos variar **b** y mantenemos **a** constante tenemos nuevamente que Q representa a una parábola cóncava (por ser $\sum x_i^2 > 0$) que alcanzará un mínimo en su vértice, cuya abcisa vale $b = \dfrac{-2(a\sum x_i - \sum x_i y_i)}{2\sum x_i^2} \Rightarrow b\sum x_i^2 = -a\sum x_i + \sum x_i y_i \Rightarrow \boxed{a\sum x_i + b\sum x_i^2 = \sum x_i y_i}$, esta es la segunda ecuación normal.

Tenemos el sistema:

$$\left.\begin{array}{l} a + b\overline{x} = \overline{y} \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{array}\right\}$$

Se tiene que:

Media de x : $\overline{x} = \dfrac{\sum x_i}{n} \Rightarrow \boxed{\sum x_i = n\overline{x}}$

Varianza de x: $\sigma_x^2 = \dfrac{\sum x_i^2}{n} - \overline{x}^2 \Rightarrow \boxed{\sum x_i^2 = n\overline{x}^2 + n\sigma_x^2}$

Media de y : $\overline{y} = \dfrac{\sum y_i}{n} \Rightarrow \boxed{\sum y_i = n\overline{y}}$

Varianza de y: $\sigma_y^2 = \dfrac{\sum y_i^2}{n} - \overline{y}^2 \Rightarrow \boxed{\sum y_i^2 = n\overline{y}^2 + n\sigma_y^2}$

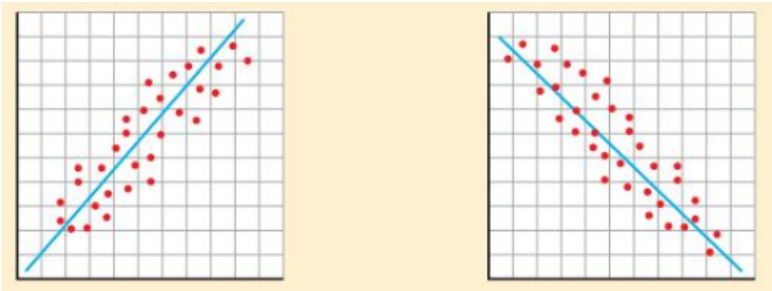
Covarianza: $\sigma_{xy} = \dfrac{\sum x_i y_i}{n} - \overline{x}\ \overline{y} \Rightarrow \boxed{\sum x_i y_i = n\sigma_{xy} + n\overline{x}\ \overline{y}}$

Volvamos al sistema de ecuaciones anterior, despejamos a en la primera ecuación $\boxed{a = \overline{y} - b\overline{x}}$, si sustituimos en la segunda ecuación obtenemos $(\overline{y} - b\overline{x})n\overline{x} + bn(\sigma_x^2 + \overline{x}^2) = n\sigma_{xy} + n\overline{x}\ \overline{y}$, con lo que $\cancel{n\overline{y}\,\overline{x}} - \cancel{bn\overline{x}^2} + bn\sigma_x^2 + \cancel{bn\overline{x}^2} = n\sigma_{xy} + \cancel{n\overline{y}\,\overline{x}} \Rightarrow b = \dfrac{\cancel{n}\sigma_{xy}}{\cancel{n}\sigma_x^2} \Rightarrow \boxed{b = \dfrac{\sigma_{xy}}{\sigma_x^2}}$. Recordemos que la recta de regresión era $y = a + bx$, que sustituyendo los valores hallados se convierte en $y = \overline{y} - \dfrac{\sigma_{xy}}{\sigma_x^2}\overline{x} + \dfrac{\sigma_{xy}}{\sigma_x^2}x$. La recta de regresión en forma punto pendiente es por tanto

$$\boxed{y - \overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x})}$$

3. Correlación

Cuando aproximamos una nube de puntos a una recta de regresión, llamamos correlación a la mayor o menor aproximación de la recta de regresión a la nube de puntos. Si hay una buena aproximación la **correlación se llama fuerte** y si la aproximación no es tan buena **la correlación se llama débil**. La correlación también nos indica la relación que existe entre las abcisas x_i y las ordenadas y_i de la nube de puntos (x_i, y_i). Esta correlación es positiva si al aumentar x_i también aumenta y_i, y es negativa si al aumentar x_i disminuye y_i



3.1. Coeficiente de determinación

Lamamos variación explicada por la recta de regresión a $VE = \sum(\hat{y}_i - \overline{y})^2$, llamamos variación total a $VT = \sum(y_i - \overline{y})^2$. Se define el coeficiente de determinación, que se representa por R^2, como $R^2 = \dfrac{VE}{VY} = \dfrac{\sum(\hat{y} - \overline{y})^2}{\sum(y_i - \overline{y})^2}$. Vamos a poner esta expresión en función de la covarianza y la varianza.

$VE = \sum(\hat{y}_i - \overline{y})^2 = \sum(\cancel{a} + bx_i - \cancel{a} - b\overline{x})^2 = \sum b^2(x_i\overline{x})^2 = nb^2\sigma_x^2$. Recordemos que $b = \dfrac{\sigma_{xy}}{\sigma_x^2}$. Entonces $VE = n\dfrac{\sigma_{xy}^2}{(\sigma_x^2)^2}\sigma_x^2 = n\dfrac{\sigma_{xy}^2}{\sigma_x^2}$. Es evidente que $VT = \sum(y_i - \overline{y})^2 = n\sigma_y^2$ y por tanto el coeficiente de determinación lo posemos escribir como $R^2 = \dfrac{\sigma_{xy}^2}{\sigma_x^2\ \sigma_y^2}$. El coeficiente de correlación lineal se representa por r, y se define como $r = \pm\dfrac{\sigma_{xy}}{\sigma_x\ \sigma_y}$ y es evidente que r=$\pm\sqrt{R^2}$. El signo de r depende de si la regresión es positiva o negativa. ¿Cómo interpretamos estos valores? Observemos que R^2 representa el porcentaje de la variación total que se explica por el modelo de la recta de regresión. Al verificarse la desigualdad $\sigma xy \leq \sigma_x\ \sigma_y$ también se cumple que $-1 \leq r \leq 1$.

3.2. ¿Cómo se explica la correlación?

i) Supongamos que r=0. Entonces $R^2 = 0$, eso quiere decir que la variación del modelo explica el 0 % de la variación total, es decir que el modelo y la nube de puntos no tienen nada que ver, no hay correlación, la variable $\hat{y}_1$ del modelo es independiente de la variable x_i

ii) Si $r = 1$ o $r = -1$ entonces $R^2 = 1$, la variación del modelo explica al 100 % la variación total, hay correlación perfecta, la variable $\hat{y}_1$ del modelo depende de la variable x_i

iii) Cuanto mayor sea r en valor absoluto, mejor explica el modelo el comportamiento de la nube de puntos.

La recta de regresión se utiliza para hacer predicciones. Veamos un ejemplo.

En la siguiente tabla se dan las notas en matemáticas y física de una clase de 25 alumnos

Matemáticas	1	4	5	6	7	8	9	10
Física	3	5	7	8	7	7	10	10
Nº Alumnos	4	4	5	6	1	2	2	1

Si un alumno obtiene una nota de 4'5 en matemáticas, ¿qué nota obtendrá en física?

Solución:

Sea $x_i \equiv matemáticas$, $y_i \equiv física$, $f_i \equiv n^o\ de\ alumnos$, $frecuencia\ absoluta$. A partir de la tabla anterior generamos la tabla siguiente:

x_i	1	4	5	6	7	8	9	10
y_i	3	5	7	8	7	7	10	10
f_i	4	4	5	6	1	2	2	1
$x_i f_i$	4	16	25	36	7	16	18	10
$y_i f_i$	12	20	35	48	7	14	10	10
$x_i^2 f_i$	1	64	125	216	49	128	162	100
$y_i^2 f_i$	36	100	245	384	49	98	200	100
$x_i y_i f_i$	12	80	175	288	49	112	180	100

Si sumamos las filas correspondientes obtenemos que:

$$\sum x_i f_i = 132 \Rightarrow \overline{x} = \frac{132}{25} = 5{,}28$$

$$\sum y_i f_i = 156 \Rightarrow \overline{y} = \frac{156}{25} = 6{,}24$$

$$\sum x_i^2 f_i = 845 \Rightarrow \sigma_x^2 = \frac{845}{25} - 5{,}28^2 = 5{,}92. \text{ Entonces } \sigma_x = 2{,}43$$

$$\sum y_i^2 f_i = 1212 \Rightarrow \sigma_y^2 = \frac{1212}{25} - 6{,}24^2 = 9{,}54. \text{ Entonces } \sigma_y = 3{,}09$$

$$\sum x_i y_i f_i = 996 \Rightarrow \sigma_{xy} = \frac{996}{25} - 5{,}28 \cdot 6{,}24 = 6{,}89.$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{6{,}89}{5{,}92} = 1{,}16 \text{ (esta es la pendiente de la recta de regresión)}$$

El coeficiente de correlación lineal vale $\frac{6{,}89}{2{,}43 \cdot 3{,}09} = 0{,}918$, como el coeficiente de correlación lineal de Pearson es muy próximo a 1, esto quiere decir que la recta de regresión describe bien la dependencia entre las variables $\mathbf{x}$ e $\mathbf{y}$. La ecuación de la recta de regresión es $y - 6{,}24 = 1{,}16(x - 5{,}28)$. Si la nota en matemáticas es de $x = 4{,}5$ la nota de física que se espera es $y - 6{,}24 = 1{,}16(4{,}5 - 5{,}28) \Rightarrow\ y = 5{,}19$, esta es la nota esperada de física.