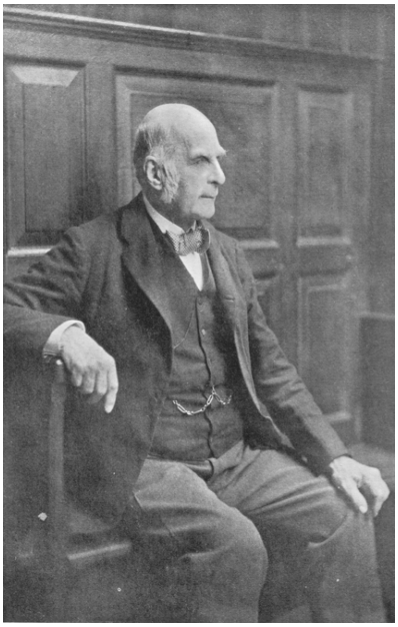
Introducción



Francis Galton
Inglés, 1822-1911

El término regresión fue introducido por Galton en su libro "*Natural inheritance*" (1889) refiriéndose a la "ley de la regresión universal":

– "Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor." (Regresión a la media)

– Su trabajo se centraba en la descripción de los rasgos físicos de los descendientes (una variable) a partir de los de sus padres (otra variable).

– Pearson (un amigo suyo) realizó un estudio con más de 1000 registros de grupos familiares observando una relación del tipo:

  • Altura del hijo = 85cm + 0,5 • altura del padre (aprox.)

  • Conclusión: los padres muy altos tienen tendencia a tener hijos que heredan parte de esta altura, aunque tienen tendencia a acercarse (*regresar*) a la media. Lo mismo puede decirse de los padres muy bajos.

Hoy en día el sentido de regresión es el de predicción de una medida basándonos en el conocimiento de otra.

Método de los mínimos cuadrados.

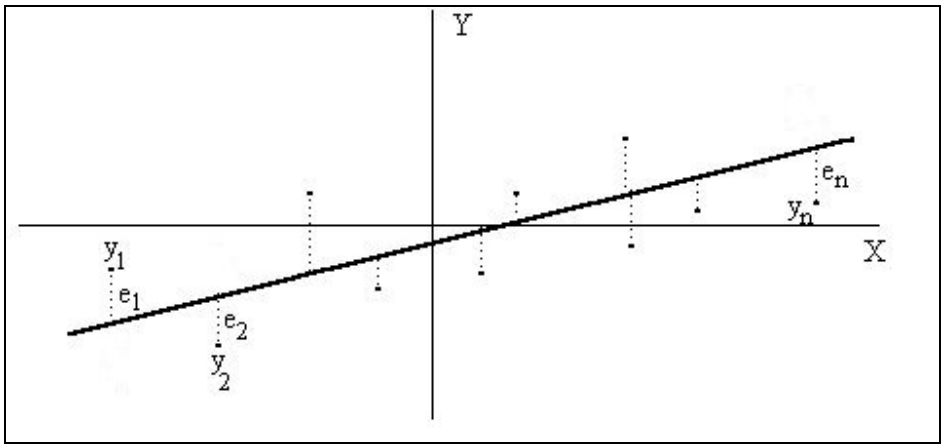
Dado un conjunto de *n* puntos

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n) \qquad (*)$$

resultantes de datos tomados en el estudio de un fenómeno particular. Se desea, como es de suponer, encontrar una función $y=f(x)$ que ajuste *adecuadamente* estos datos experimentales. Como ya se ha comentado, en primer lugar se grafica la *nube de puntos* (*), y de acuerdo a forma que asume esta nube se elige la función con la cual se modelará la situación. La función elegida depende, en general, de algunos parámetros. Para encontrar los parámetros del mejor modelo se usará el *método de los mínimos cuadrados*.

Ajuste lineal por el método de los mínimos cuadrados.

Supongamos que al graficar los puntos (*) en un sistema de coordenadas resultan tener un comportamiento *aproximadamente* lineal. ¿Como encontrar una recta $y = mx + b$ que *mejor* ajuste estos valores?. Matemáticamente, esta recta se encuentra *minimizando los cuadrados de las distancias verticales entre los n puntos y la recta buscada*. Este procedimiento es conocido con el nombre *método de los mínimos cuadrados*.



Para encontrar la recta buscada se debe minimizar:

$$E = (mx_1 + b - y_1)^2 + (mx_2 + b - y_2)^2 + \cdots + (mx_n + b - y_n)^2 = \sum_{i=1}^{n} (mx_i + b - y_i)^2$$

Los parámetros m y b buscados, deben satisfacer que $\frac{\partial E}{\partial m}=0$ y $\frac{\partial E}{\partial b}=0$. Calculando estas derivadas parciales, se obtiene el siguiente sistema de ecuaciones lineales:

$$\begin{cases} \sum_{i=1}^{n}\left(mx_i+b-y_i\right)x_i=0 \\ \sum_{i=1}^{n}\left(mx_i+b-y_i\right)=0 \end{cases}$$

es decir:

$$\begin{cases} \left(\sum x_i^{\,2}\right)m+\left(\sum x_i\right)b=\sum x_i y_i \\ \left(\sum x_i\right)m+nb=\sum y_i \end{cases} \quad (**)$$

resolviendo este sistema de ecuaciones, se obtiene:

$$m=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2} \quad \text{y} \quad b=\frac{\sum_{i=1}^{n}x_i^{\,2}\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i y_i}{n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2}$$

que son los parámetros del ajuste lineal buscado.

Ejemplo

Implementemos este método para buscar la recta que mejor ajusta, según el criterio de los mínimos cuadrados, los siguientes puntos:

x	1	2	3
y	1	3	4

Tabla 1

(**) Se han omitido, por comodidad, los índices en las sumatorias.

Para buscar los valores de m y b ordenemos los elementos necesarios para calcularlos en la siguiente tabla:

	x	**y**	**x^2**	**xy**
	1	1	1	1
	2	3	4	6
	3	4	9	12
$\sum$	6	8	14	19

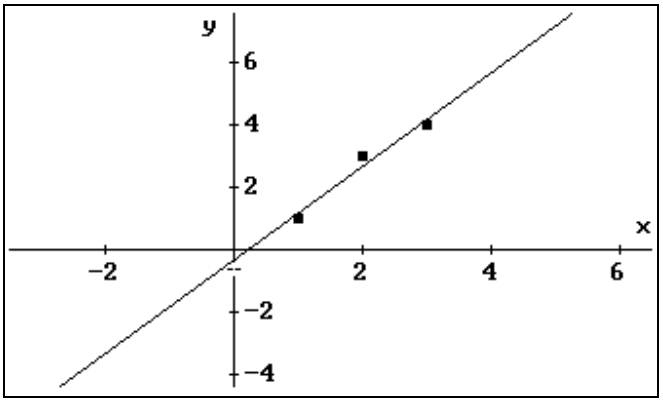
Luego,

$$m = \frac{3\cdot 19 - 6\cdot 8}{3\cdot 14 - 36} = 1.5 \qquad \text{y} \qquad b = \frac{14\cdot 8 - 6\cdot 19}{3\cdot 14 - 36} \approx -0.33\,.$$

Por lo tanto, la recta que mejor ajusta, por el método de los mínimos cuadrados, los valores de la tabla 1, es

$$y = 1.5x - 0.33\,.$$

En el siguiente gráfico se indican los puntos involucrados junto a la recta encontrada.



¿Qué tan bien representa a los datos experimentales la recta encontrada?

Existe un parámetro, llamado coeficiente de correlación[1], que permite estimar el grado de asociación lineal entre las variables en estudio. Se define por

$$R^2 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left(n\sum x_i{}^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i{}^2 - \left(\sum y_i\right)^2\right)}}$$

o equivalentemente:

$$R^2 = \frac{\Sigma\left(\hat{y}_i - \bar{y}\right)^2}{\Sigma\left(y_i - \bar{y}\right)^2}$$

donde, $\bar{y} = \dfrac{\sum y_i}{n}$ e $\hat{y} =$ valor estimado por el ajuste.

Es posible verificar que:

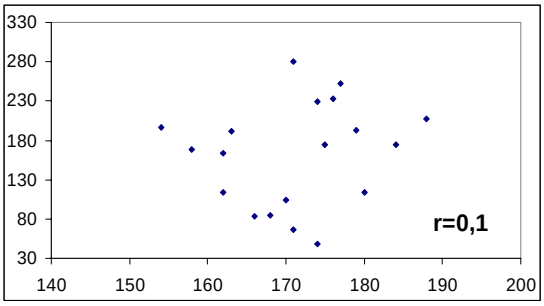
- R varía entre -1 y 1.
- Mientras más cerca de los extremos del rango de variación, es decir de -1 o 1, se encuentre R, la recta de regresión lineal mejor ajusta los datos del problema estudiado.
- Habitualmente se exigen valores de r superiores (en módulo) a 0,75 para poder decir que existe una fuerte dependencia lineal entre las variables en estudio. En general:

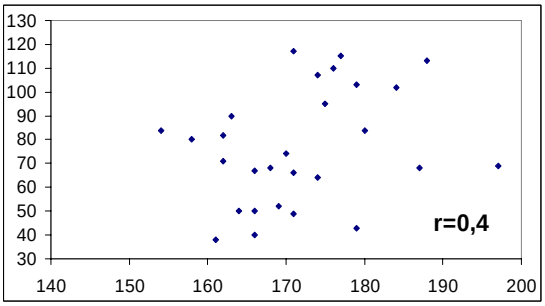
|R|	**Correlación**
0.0 a 0.2	muy débil, despreciable
0.2 a 0.4:	débil. bajo
0.4 a 0.7:	moderada
0.7 a 0.9	fuerte, alto, importante
0.9 a 1.0	muy fuerte, muy alto

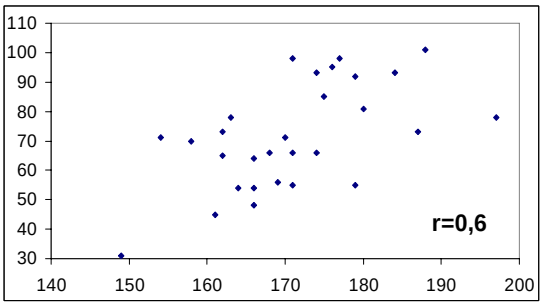
Observación: La correlación entre los valores de dos variables es un hecho. El que lo consideremos satisfactorio o no, depende de la interpretación. Otro problema que representa la correlación es cuando se pregunta si una variable, de algún modo causa o determina a la otra. La correlación no implica causalidad. Si las variables X e Y están correlacionadas, esto puede ser por que X causa a Y, o porque Y causa a X o porque alguna otra variable afecta tanto a X como Y, o por una combinación de todas estas razones; o puede ser que la relación sea una *coincidencia*.

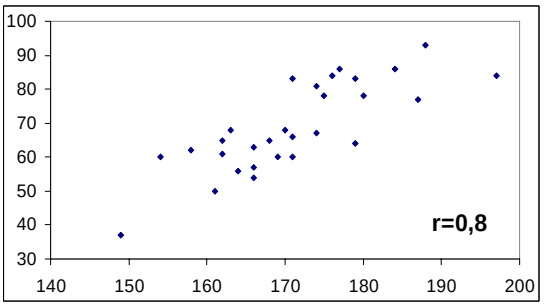
[1] Llamado también coeficiente de correlación de Pearson

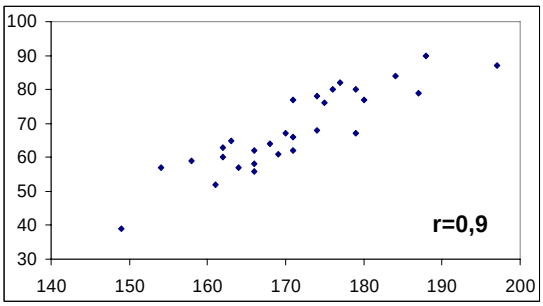
Ejemplos gráficos

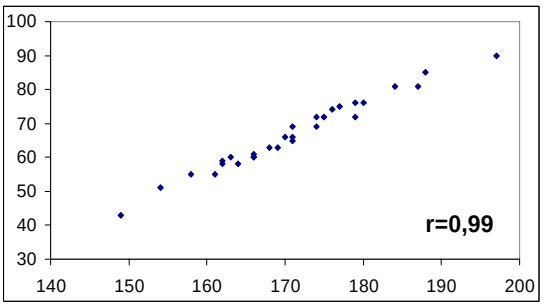


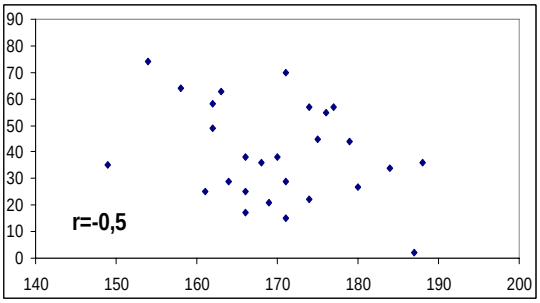


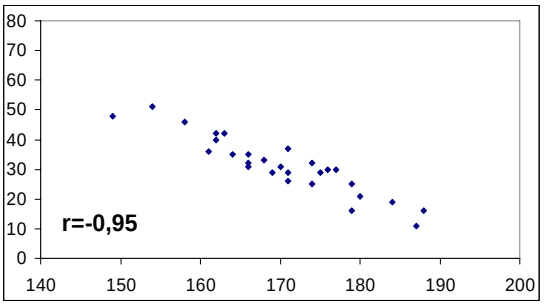












Actividad

Calcular el coeficiente de correlación de la recta encontrada en el ejemplo 1.

Ajuste cuadrático por el método de los mínimos cuadrados.

Si la nube de puntos (*) provenientes de una situación concreta que se desea modelar, *insinúan* una forma parabólica, se tiene que la situación en estudio se modela por una función cuadrática. Para encontrar esta función cuadrática, es posible también usar el método de los mínimos cuadrados.

Sea $y = a + bx + cx^2$ la cuadrática buscada que ajusta los *n* puntos (*). Luego se debe minimizar

$$E = \sum \left(a + bx_i + cx_i^2 - y_i \right)^2$$

Calculando las derivadas parciales de E (con respecto a *a*, *b* y *c*) e igualándolas a 0 se obtiene el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} \left(\sum x_i^2\right)a + \left(\sum x_i^3\right)b + \left(\sum x_i^4\right)c &= \sum x_i^2 y_i \\ \left(\sum x_i\right)a + \left(\sum x_i^2\right)b + \left(\sum x_i^3\right)c &= \sum x_i\, y_i \\ na + \left(\sum x_i\right)b + \left(\sum x_i^2\right)c &= \sum y_i \end{aligned}$$

Resolviendo este sistema, se obtienen los parámetros de la función cuadrática buscada.

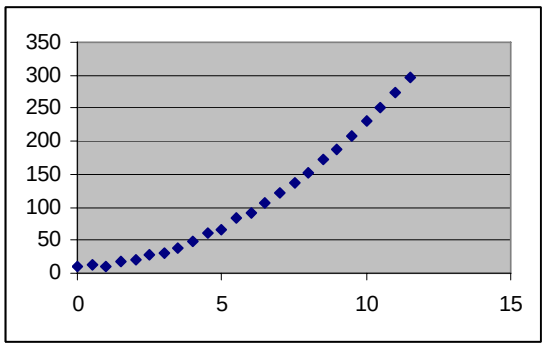
Ejemplo

En determinado proceso se realizaron una serie de 24 mediciones, que luego al graficarse se determinó que es de naturaleza cuadrática. Se desea encontrar los parámetros del polinomio de segundo grado, que mejor se ajusta a esa serie de datos, y cuál es el valor de la variable dependiente, cuando el valor de la variable independiente es de 20.

La tabla con los datos medidos es la siguiente:

X	Y
0	10,08
0,5	12,03
1	11,38
1,5	18,81
2	20,53
2,5	28,50
3	31,38
3,5	38,40
4	48,39
4,5	60,60
5	66,66
5,5	82,61
6	91,37
6,5	105,44
7	122,53
7,5	137,77
8	152,74
8,5	172,65
9	188,84
9,5	207,77
10	230,94
10,5	251,35
11	274,07
11,5	295,95

Tabla 2



Nube de puntos de la Tabla 2

Ahora, teniendo en cuenta el sistema de ecuaciones que se dedujo anteriormente, se sabe que se tienen que encontrar los valores de la suma de x, la suma de x^2, de x^3, x^4, de y, xy, x^2*y y n=24. Como es de suponer, la siguiente tabla se ha generado en una planilla excel.

	x	**y**	**x^2**	**x^3**	**x^4**	**xy**	**x^2 y**
	0	10,08	0,00	0,00	0,00	0,00	0,00
	0,5	12,03	0,25	0,13	0,06	6,01	3,01
	1	11,38	1,00	1,00	1,00	11,38	11,38
	1,5	18,81	2,25	3,38	5,06	28,21	42,31
	2	20,53	4,00	8,00	16,00	41,06	82,13
	2,5	28,50	6,25	15,63	39,06	71,24	178,11
	3	31,38	9,00	27,00	81,00	94,14	282,41
	3,5	38,40	12,25	42,88	150,06	134,39	470,36
	4	48,39	16,00	64,00	256,00	193,56	774,26
	4,5	60,60	20,25	91,13	410,06	272,68	1227,08
	5	66,66	25,00	125,00	625,00	333,31	1666,55
	5,5	82,61	30,25	166,38	915,06	454,37	2499,02
	6	91,37	36,00	216,00	1296,00	548,23	3289,38
	6,5	105,44	42,25	274,63	1785,06	685,39	4455,05
	7	122,53	49,00	343,00	2401,00	857,74	6004,20
	7,5	137,77	56,25	421,88	3164,06	1033,24	7749,32
	8	152,74	64,00	512,00	4096,00	1221,90	9775,23
	8,5	172,65	72,25	614,13	5220,06	1467,54	12474,08
	9	188,84	81,00	729,00	6561,00	1699,55	15295,92
	9,5	207,77	90,25	857,38	8145,06	1973,80	18751,13
	10	230,94	100,00	1000,00	10000,00	2309,40	23093,97
	10,5	251,35	110,25	1157,63	12155,06	2639,18	27711,38
	11	274,07	121,00	1331,00	14641,00	3014,81	33162,86
	11,5	295,95	132,25	1520,88	17490,06	3403,37	39138,76
∑	**138**	**2660,8**	**1081**	**9522**	**89452,75**	**22494,51**	**208137,88**

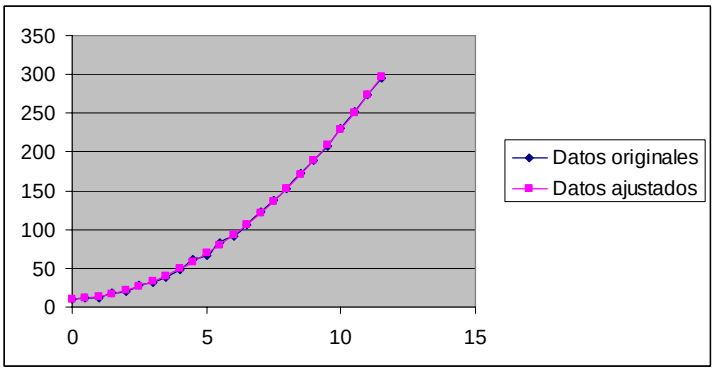
luego, el sistema que se debe resolver es:

$$\begin{array}{rcrcrcr} 24\,a & + & 138\,b & + & 1081\,c & = & 2660{,}8 \\ 138\,a & + & 1081\,b & + & 9522\,c & = & 22495 \\ 1081\,a & + & 9522\,b & + & 89453\,c & = & 208138 \end{array}$$

Resolviendo este sistema se obtiene que: ***a* = 9.6**, ***b* = 1.76** y ***c* = 2.02**. Luego la función cuadrática que mejor ajusta, por el método de los mínimos cuadrados, la serie de datos en estudio es:

$$y = 9.6 + 1.76x + 2.02x^2$$

En el siguiente gráfico, se muestran los datos originales junto al gráfico de la función cuadrática encontrada:




Como es de suponer, el método de los mínimos cuadrados permite determinar un polinomio de grado m de ajuste la serie de datos experimentales (*). En este caso, la función buscada es

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$$

con $n \geq m+1$. El polinomio buscado se encuentra minimizando la función (de m variables):

$$\Pi = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} \left[y_i - (a_0 + a_1 x_i + a_2 x_i^{\,2} + \ldots + a_m x_i^{\,m}) \right]^2 = \min$$

Luego,

$$\frac{\partial \Pi}{\partial a_0} = 2\sum_{i=1}^{n} \left[y_i - (a_0 + a_1 x_i + a_2 x_i^{\,2} + \ldots + a_m x^m) \right] = 0$$

$$\frac{\partial \Pi}{\partial a_1} = 2\sum_{i=1}^{n} x_i \left[y_i - (a_0 + a_1 x_i + a_2 x_i^{\,2} + \ldots + a_m x^m) \right] = 0$$

$$\frac{\partial \Pi}{\partial a_2} = 2\sum_{i=1}^{n} x_i^{\,2} \left[y_i - (a_0 + a_1 x_i + a_2 x_i^{\,2} + \ldots + a_m x^m) \right] = 0$$

$$\vdots$$

$$\vdots$$

$$\frac{\partial \Pi}{\partial a_m} = 2\sum_{i=1}^{n} x_i^{\,m} \left[y_i - (a_0 + a_1 x_i + a_2 x_i^{\,2} + \ldots + a_m x^m) \right] = 0$$

De donde, se obtiene el siguiente sistema de *m* ecuaciones lineales con *m* incógnitas:

$$\sum_{i=1}^{n} y_i = a_0 \sum_{i=1}^{n} 1 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^{\,2} + \ldots + a_m \sum_{i=1}^{n} x_i^{\,m}$$

$$\sum_{i=1}^{n} x_i y_i = a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^{\,2} + a_2 \sum_{i=1}^{n} x_i^{\,3} + \ldots + a_m \sum_{i=1}^{n} x_i^{\,m+1}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a_0 \sum_{i=1}^{n} x_i^{\,2} + a_1 \sum_{i=1}^{n} x_i^{\,3} + a_2 \sum_{i=1}^{n} x_i^{\,4} + \ldots + a_m \sum_{i=1}^{n} x_i^{\,m+2}$$

$$\vdots$$

$$\vdots$$

$$\sum_{i=1}^{n} x_i^{\,m} y_i = a_0 \sum_{i=1}^{n} x_i^{\,m} + a_1 \sum_{i=1}^{n} x_i^{\,m+1} + a_2 \sum_{i=1}^{n} x_i^{\,m+2} + \ldots + a_m \sum_{i=1}^{n} x_i^{\,2m}$$

Resolviendo este sistema, se encuentran los *m*+1 parámetros del polinomio (de grado *m*) buscado.