Regresión

josé a. mañas

8.2.2017

1 Introducción

El objetivo de las técnicas de regresión es identificar una función que permita estimar una variable Y en función de la otra X. Es decir, averiguar una función

$$y = f(x)$$

que represente lo mejor posible la relación entre valores X e Y permitiéndonos inferir un valor a partir del otro.

2 Definiciones

Dado un conjunto de pares de datos experimentales <x, y>, se definen varios estadísticos:

Definiciones	
valor medio de X	$\bar{x} = \frac{\sum x_i}{n}$
valor medio de Y	$\bar{y} = \frac{\sum y_i}{n}$
desviación típica de X	$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
desviación típica de Y	$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$
covarianza XY	$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
índice de correlación (Pearson)	$r = \frac{s_{xy}}{s_x s_y}$

El valor del índice de correlación varía en el intervalo [-1,1]:

- Si r = 1, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
- Si 0 < r < 1, existe una correlación positiva.

- Si r = 0, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.
- Si -1 < r < 0, existe una correlación negativa.
- Si r = -1, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en proporción constante.

Como regla aproximada,

- correlación fuerte: $|r| > 0.8$
- correlación débil: $|r| < 0.5$

aunque a menudo lo mejor es representar los datos gráficamente para verlo.

3 Mínimos cuadrados

Mínimos cuadrados es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados: variable independiente, variable dependiente, y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de *mínimo error cuadrático*.

En su forma más simple, intenta minimizar la suma de cuadrados de las diferencias *en las ordenadas* entre los puntos generados por la función elegida y los correspondientes valores en los datos.

Desde un punto de vista estadístico, un requisito implícito para que funcione el método de mínimos cuadrados es que los errores de cada medida estén distribuidos de forma aleatoria. También es importante que los datos a procesar estén bien escogidos, para que permitan visibilidad en las variables que han de ser resueltas.

Formalmente, dado un conjunto de puntos experimentales <x, y> se trata de encontrar una función y= f(x) tal que minimice la suma de los cuadrados de las diferencias entre los valores medidos y los calculados usando la fórmula; es decir, minimizar

$$\sum (y - f(x))^2$$

4 Regresión lineal

Buscamos una relación lineal entre x e y; es decir

$$y = ax + b$$

a y b se calculan como

$$a = \frac{s_{xy}}{s_x^2}$$

$$b = \bar{y} - a\bar{x}$$

Para estimar cómo de buena es nuestra estimación, se usa el coeficiente de determinación r^2, que es el cuadrado del coeficiente de correlación de Pearson.

r^2 es útil porque nos da la proporción en que la varianza de la variable Y es predecible en función de la variable X. En otras palabras, es la proporción de la variabilidad de Y que se puede explicar como consecuencia de la variación de X.

Una regresión lineal perfecta es la que permite predecir Y al 100% conocido X; es decir, la que tiene $r^2 = 1$.

Ejemplo. Si $r^2 = 0{,}85$, diremos que el 85% de la varianza de Y es explicable. Y viceversa, el 15% es inexplicable (es decir, será consecuencia de otros factores aparte de X).

Gráficamente, $r^2 = 1$ significa que, gráficamente, la línea de regresión pasa exactamente por todos los puntos, mientras que un r^2 muy bajo indica que los puntos no se ajustan muy bien a la línea.

5 Regresión no lineal

Dada una serie de puntos <x, y> que no se ajustan a una relación lineal, una forma sencilla de tratar el problema es transformar las variables para que se ajusten a una relación lineal.

5.1 Logarítmica

Si sospechamos que los puntos están relacionados por una función del tipo

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

podemos hacer la transformación

$$y' = y$$

$$x' = \log(x)$$

y resolver el problema de una regresión lineal

$$y' = a\,x' + b$$

5.2 Potencial

Si sospechamos que los puntos están relacionados por una función del tipo

$$y = bx^a$$

podemos hacer la transformación

$$y' = \log(y)$$

$$x' = \log(x)$$

$$b' = \log(b)$$

y resolver el problema de una regresión lineal

$$y' = a\,x' + b'$$

5.3 Exponencial

Si sospechamos que los puntos están relacionados por una función del tipo

$$y = ba^x$$

podemos hacer la transformación

$$y' = \log(y)$$

$$x' = x$$

$$a' = \log a$$

$$b' = \log b$$

y resolver el problema de una regresión lineal

$$y' = a'\,x' + b'$$

6 Ejemplos

Usando la aplicación "correlator" que puede encontrar en la web de la asignatura.

6.1 Regresión lineal

Sean los datos experimentales

X	Y
30	200
50	400
50	800
60	1.200
60	900

Aplicamos las fórmulas

a	28,33
b	-716,67
r^2	0,75

Es decir, que podemos hacer una aproximación no muy buena usando

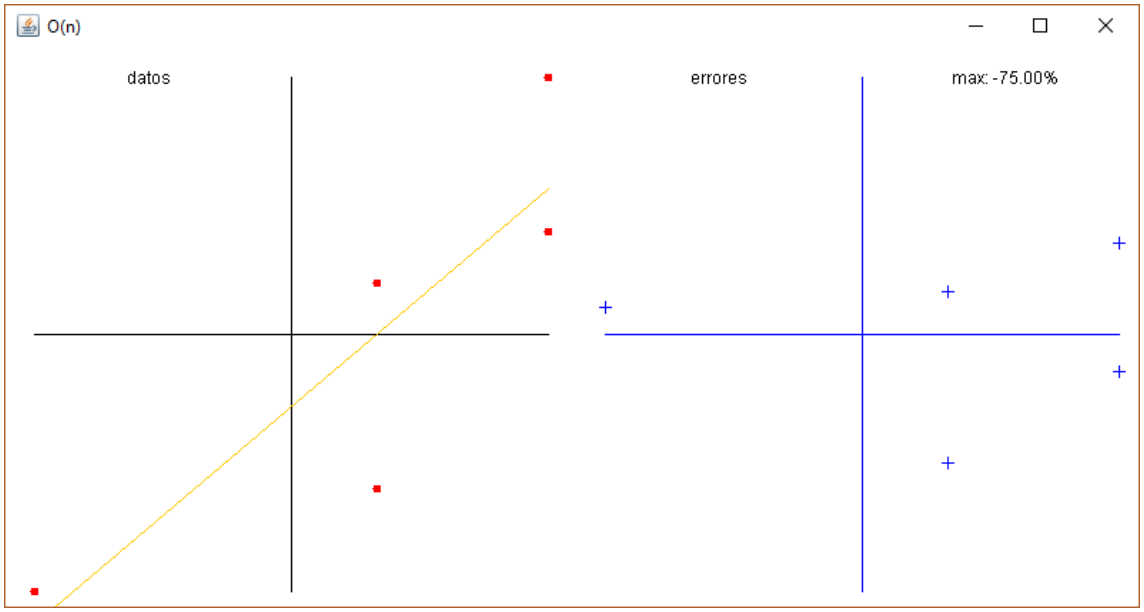
$$y = 28{,}33x - 716{,}67$$

Correlator (2.3.2016)

30	200
50	400
50	800
60	1200
60	900

complejidad	a	b	r^2
O(log(n))	1.2e+03	-3.9e+03	0.71
O(n)	28	-7.2e+02	0.75
O(n log(n))	5.9	-4.7e+02	0.76
$O(n^a)$	2.3	0.065	0.85
$O(a^n)$	1.1	38	0.86

Gráficamente



6.2 Regresión lineal

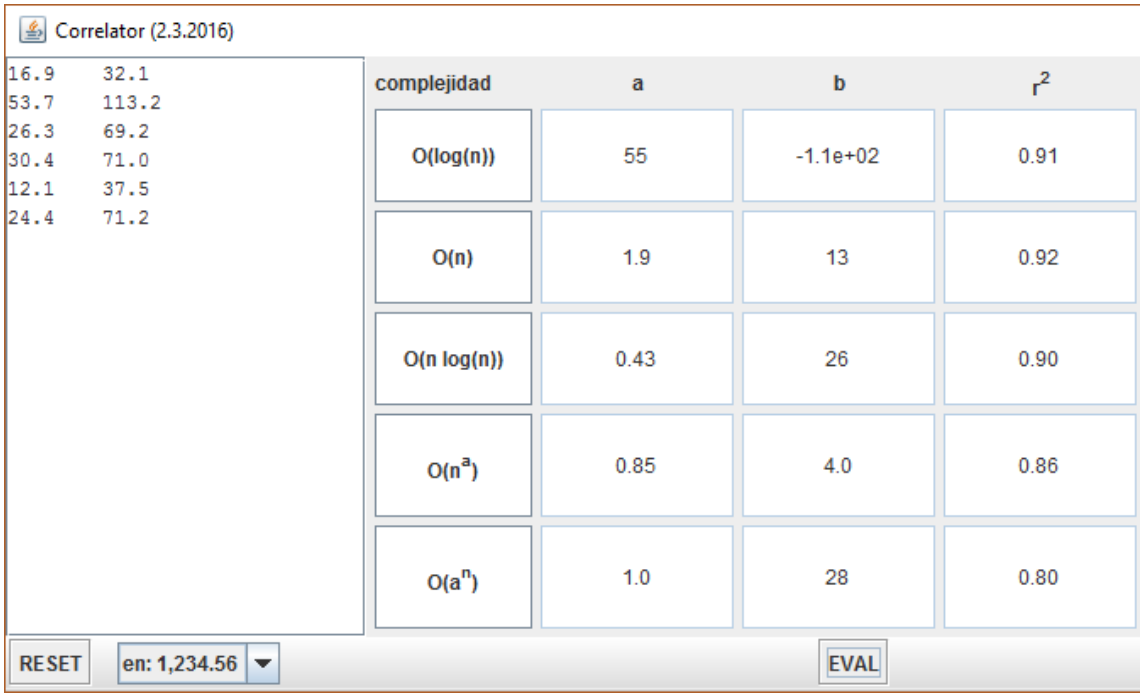
Datos

X	Y
16.9	32.1
53.7	113.2
26.3	69.2
30.4	71.0
12.1	37.5

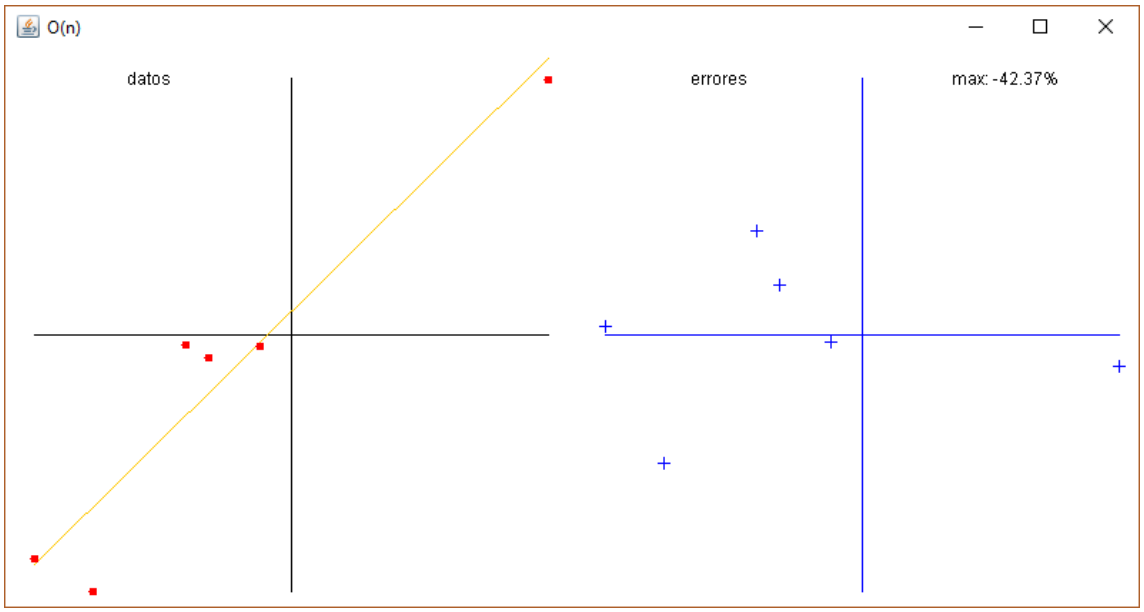
24.4	71.2

Recta de regresión

a	1,92
b	13,24
r^2	0,92



Gráficas

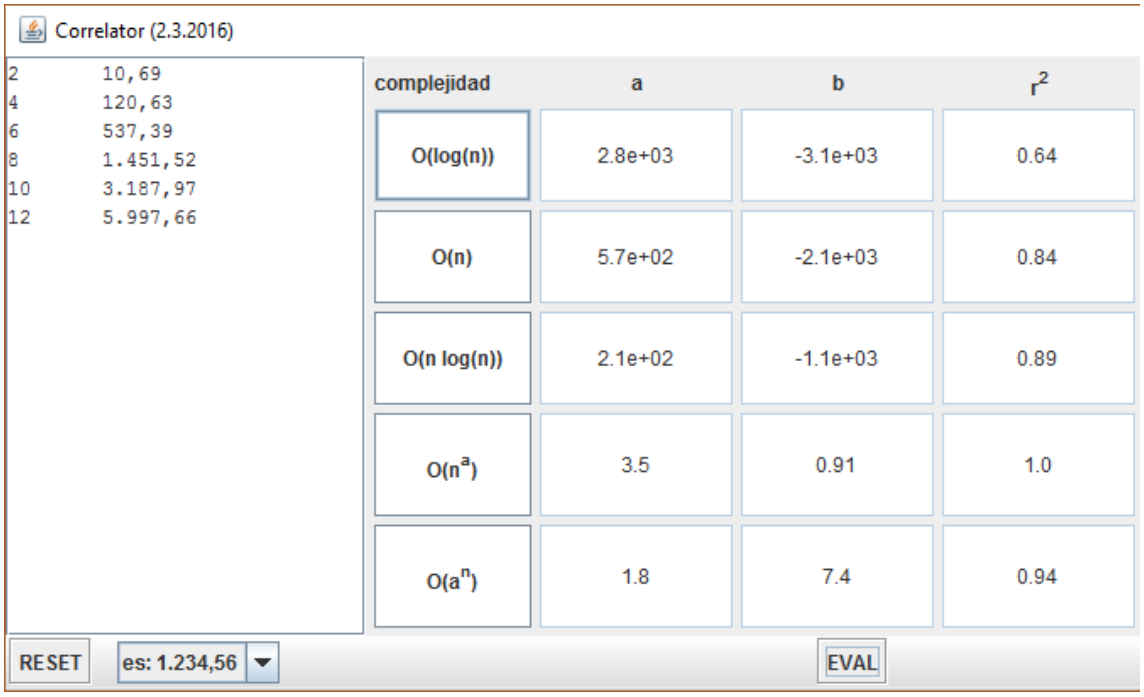


6.3 Ajuste potencial

Datos experimentales

X	Y
2	10,69
4	120,63
6	537,39
8	1.451,52
10	3.187,97
12	5.997,66

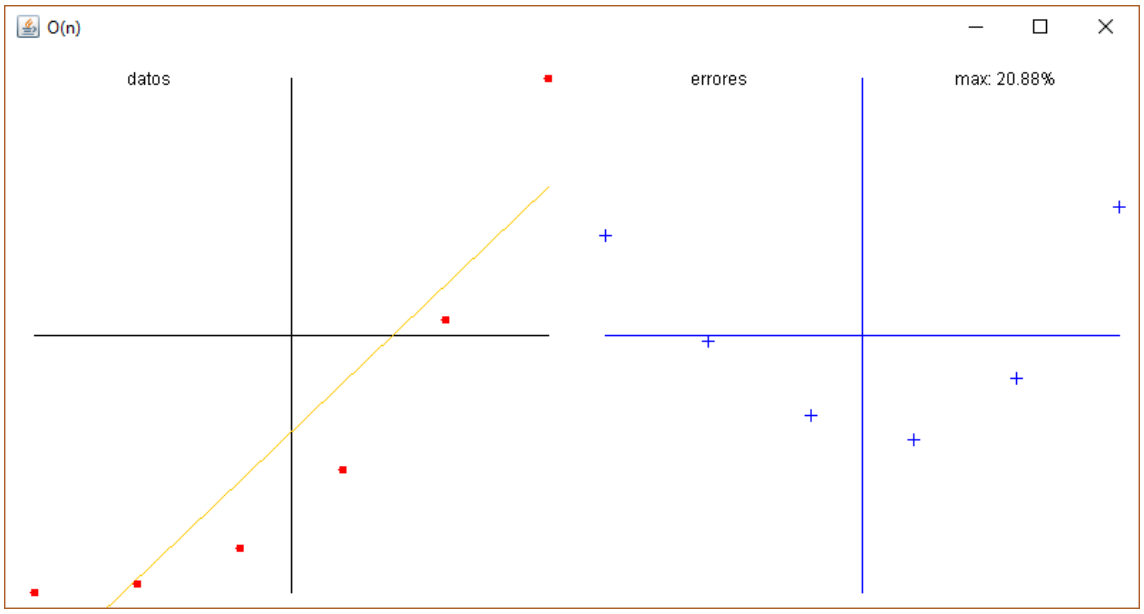
Intentamos varias fórmulas para aproximarnos



Para una regresión lineal, la recta de regresión sería:

a	572,16
b	-2.120,79
r^2	0,84

Gráficas



Parece evidente que

1. una línea recta no es una buena forma de predecir valores
2. los residuos siguen un patrón

Sospechamos que sea una relación potencial. Vamos a comprobarlo.

Hacemos el cambio de variable

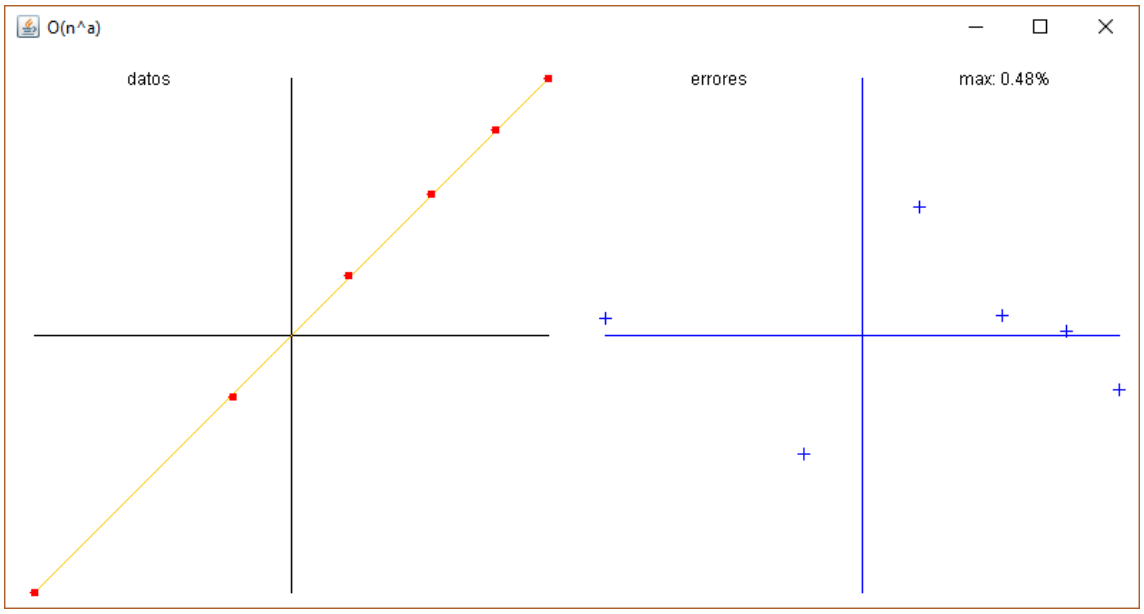
Datos experimentales

X	Y	X' = log(X)	Y' = log(Y)
2	10,69	0,69	2,37
4	120,63	1,39	4,79
6	537,39	1,79	6,29
8	1.451,52	2,08	7,28
10	3.187,97	2,30	8,07
12	5.997,66	2,48	8,70

Intentamos una regresión lineal. Recta de regresión

a'	3,54	a	3,54
b'	-0,0899	b	0,91
r^2	1.00		

Gráficas



Aunque el valor r2 ya es muy significativo de que hemos acertado en la predicción, la gráfica es contundente

1. la predicción es excelente
2. los residuos no siguen un patrón: son aleatorios

Podemos concluir que una buena aproximación es

$$y = 0.9x^{3.5}$$