Artículo
Revista digital Matemática, Educación e Internet (http://www.tec-digital.itcr.ac.cr/revistamatematica/). Vol 13, No 1. Agosto – Febrero 2013.

Análisis de regresión para la población de Costa Rica.

Luis A.Acuña P.
lacuna@itcr.ac.cr
Escuela de Matemática
Instituto Tecnológico de Costa Rica

Resumen. Breve introducción al análisis de regresión y a la transformación de algunos problemas no lineales en problemas lineales. Aplicación al caso de la población de Costa Rica como función del tiempo.

Palabras clave: Regresión, regresión lineal, regresión no lineal, predicción, crecimiento exponencial.

Abstract. Short introduction to regression analysis and the transformation of some non-linear problems to linear problemas. Application to the case of Costa Rica's population as a function of time.

KeyWords: Regression, linear regression, non-linear regression, prediction, exponential growth.

1.1 Qué es la regresión

El análisis de regresión es una técnica estadística que permite encontrar una ecuación que aproxime una variable como función de otras. Típicamente, las variables son atributos de los individuos en una población, y el análisis trabaja a partir de los valores de los atributos para alguna muestra de individuos. La variable que se escribe como función de las otras se llama *resultado*, y las otras son los *predictores*. La *regresión simple* se usa cuando hay un solo predictor.

Como ejemplo de esto, al relacionar la edad x en años con la estatura y en centímetros para niños menores de doce años, se busca una función $y = f(x)$. Si además la función buscada es lineal, $y = a + bx$, entonces se habla de *regresión lineal simple*.

Uno de los usos más comunes de la regresión es el de predecir el valor de y para un valor de x que no esté en la muestra. Por ejemplo, suponga que a partir de una muestra de niños con edades respectivas 3, 5, 6, 8, 9 y 11, en años, se ha encontrado la ecuación $y = 82.6 + 5.8x$ para la estatura en centímetros como función de la edad. Entonces se puede usar esa ecuación para predecir la estatura de un niño de 12 años: $x = 12$ resulta en $y = 82.6 + 5.8(12) \approx 152\,\text{cm}$, y esa es la estatura estimada a los doce años. El análisis de regresión lineal simple ha sido estudiado profundamente y sus mayores problemas ya están resueltos. Incluso muchas calculadoras de bolsillo pueden calcular los coeficientes a y b en la ecuación $y = a + bx$, a partir de algunos datos muestrales.

Cuando la regresión simple no es lineal, se habla de *regresión no lineal simple*, y este no es un problema que esté completamente resuelto. Para algunos casos particulares, sin embargo, existen técnicas para transformar un problema no lineal en uno lineal, en el que se puedan aplicar los resultados existentes de la regresión lineal. En las siguientes secciones se darán dos ejemplos de esto. Si el resultado y es función de varios predictores, entonces el problema es de *regresión múltiple*, que también puede ser lineal o no lineal. En regresión lineal múltiple, el resultado y se escribe como función lineal de los predictores $x_1, x_2, \ldots, x_n$, en la forma $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$.

El problema de regresión lineal múltiple también es bien conocido y presenta pocas dificultades. En particular, la regresión polinomial, en la que se busca escribir un resultado y como función polinomial de uno o varios predictores, puede transformarse fácilmente a uno de regresión lineal múltiple. Como ejemplo concreto, considere el problema de encontrar una ecuación cuadrática $y = at^2 + bt + c$ que exprese el resultado y en términos del predictor t. Si se definen dos nuevas variables $x_1 = t$ y $x_2 = t^2$, entonces la ecuación se convierte en $y = ax_2 + bx_1 + c$, que tiene la forma usual en regresión lineal múltiple.

1.2 Un ejemplo: Temperatura de agua enfriándose

La siguiente tabla muestra la temperatura, en grados centígrados, de agua en un recipiente mientras se enfría durante varios minutos ("Min" es el número de minutos transcurridos).

Min	Grados	Min	Grados	Min	Grados	Min	Grados
0.00	97.0	3.30	79.0	11.28	56.0	18.25	46.5
0.43	95.0	4.43	74.0	13.18	53.0	21.55	43.5
1.10	90.0	6.27	68.0	15.00	50.5	24.72	41.0
2.42	83.0	8.88	61.0	16.35	49.0	34.55	35.5

Tabla 1.1

Se denotará con x al tiempo en minutos y con y a la temperatura. En el siguiente gráfico se observa que la relación entre las variables x y y es aparentemente exponencial (con base menor que 1), pero trasladada hacia arriba. En efecto, es de esperar que conforme $x \to \infty$, el valor límite de y no será 0 como en una exponencial decreciente, sino que la temperatura límite convergerá a la temperatura ambiente.

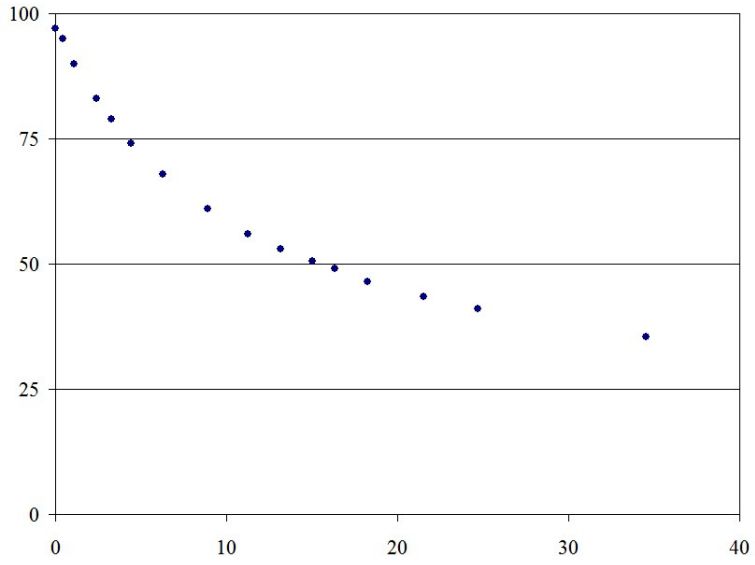

Figura 1.1 Temperatura como función del tiempo

Si se denota con TA esa temperatura ambiente, entonces puede conjeturarse que la ecuación que expresa y como función de x tiene la forma

$$y = ab^x + TA$$

donde a y b son constantes por determinar. La ecuación anterior puede convertirse en lineal de la siguiente manera:

Derechos Reservados © 2012 Revista digital Matemática, Educación e Internet (www.tec-digital.itcr.ac.cr/revistamatematica/)

$$\begin{aligned} y &= ab^x + TA \\ y - TA &= ab^x \\ \ln(y - TA) &= \ln(ab^x) \\ &= \ln a + x \ln b \\ y_1 &= a_1 + b_1 x \end{aligned}$$

donde $y_1 = \ln(y - TA)$, $a_1 = \ln a$ y $b_1 = \ln b$.

Luego de un poco de prueba y error[1] se encuentra que una buena estimación para la temperatura ambiente es $TA = 31.5$. Entonces se obtiene una nueva tabla de valores para x (que sigue siendo el número de minutos) y $y_1 = \ln(y - 31.5)$:

x	y_1	x	y_1	x	y_1	x	y_1
0.00	4.1821	3.30	3.8607	11.28	3.1987	18.25	2.7081
0.43	4.1510	4.43	3.7495	13.18	3.0681	21.55	2.4849
1.10	4.0690	6.27	3.5973	15.00	2.9444	24.72	2.2513
2.42	3.9416	8.88	3.3844	16.35	2.8622	34.55	1.3863

Tabla 1.2

Al graficar esos puntos se nota que ellos son casi colineales, lo que significa que la regresión lineal sí dará una aproximación muy cercana.

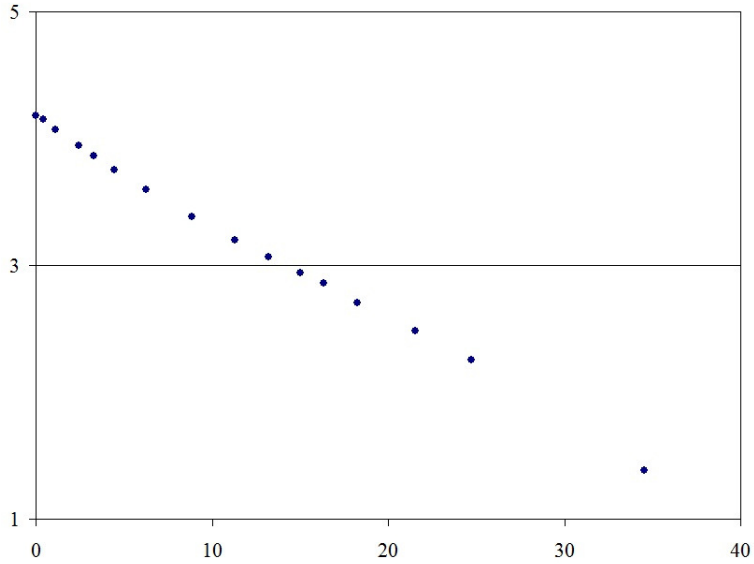

Figura 1.2 $y_1 = \ln(y - 31.5)$ como función de x

De hecho, el análisis de regresión lineal para y_1 como función de x resulta en los coeficientes $a_1 = 4.13295$ y $b = -0.078626$. Recordando que $a_1 = \ln a$ y que $b_1 = \ln b$, se despeja

$$a = \mathrm{e}^{a_1} = 62.3619 \quad \text{y} \quad b = \mathrm{e}^{b_1} = 0.924385$$

Finalmente, la ecuación $y = ab^x + TA$ se convierte en

$$\text{Temperatura} = 62.3619 \cdot 0.924385^{\text{Minutos}} + 31.5$$

[1]Se calcula el coeficiente de correlación lineal entre x y y_1 para varios valores de TA, buscando alguno que dé un coeficiente muy cercano a 1... o más bien a -1, ya que la relación es decreciente.

Al graficar los puntos en la figura 1.1 junto con esta ecuación se comprueba que efectivamente la ecuación describe las observaciones muy precisamente.

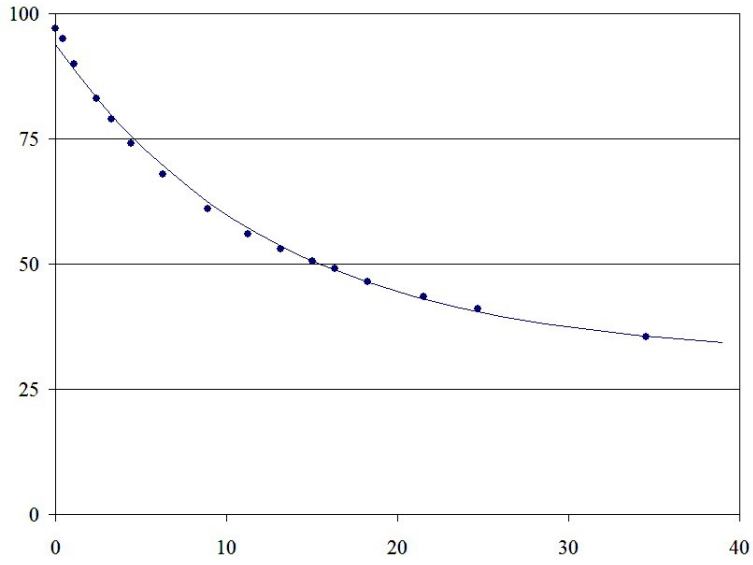

Figura 1.3 Regresión para la temperatura como función del tiempo

1.3 Evolución de la población de Costa Rica

En la figura 1.4 se ve la evolución de la población de Costa Rica, entre los años 1522 y 2000, según datos del Instituto Nacional de Estadística y Censos.

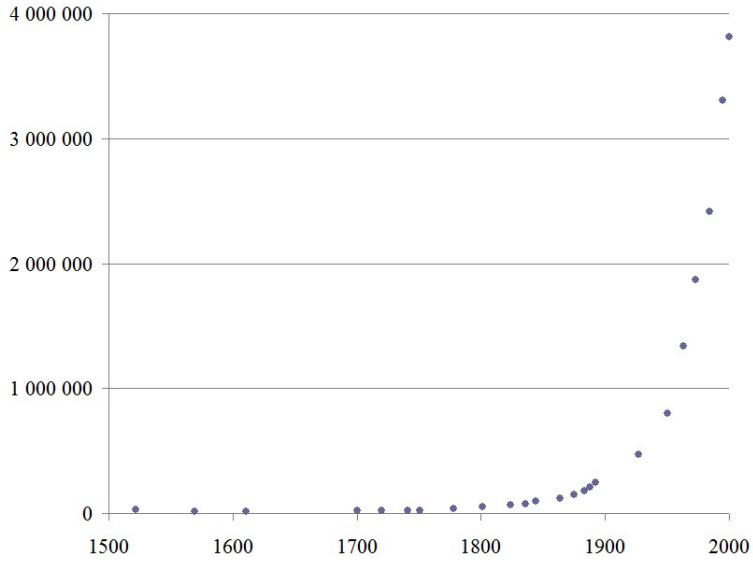

Figura 1.4 Población de Costa Rica como función del año

Derechos Reservados © 2012 Revista digital Matemática, Educación e Internet (www.tec-digital.itcr.ac.cr/revistamatematica/)

La fuente de datos para ese gráfico es la siguiente tabla.

Año	Población	Año	Población	Año	Población
1522	27 200	1801	52 591	1892	243 205
1569	17 479	1824	65 393	1927	471 524
1611	15 538	1836	78 365	1950	800 875
1700	19 293	1844	93 871	1963	1 336 274
1720	19 437	1864	120 499	1973	1 871 780
1741	24 126	1875	153 250	1984	2 416 809
1751	24 022	1883	182 073	1995	3 301 210
1778	34 212	1888	205 731	2000	3 810 179

Tabla 1.3

En el gráfico es claro que la relación entre población y tiempo no es lineal. Más bien parece exponencial, y entonces puede plantearse una ecuación de la forma

$$y = ab^t$$

donde y es la población y t el año.

Si la ecuación propuesta es correcta, entonces al tomar logaritmo natural en ambos lados se obtiene la relación lineal

$$\ln y = \ln(ab^t) = \ln a + t \ln b$$

o bien

$$y_1 = a_1 + b_1 t$$

donde $y_1 = \ln y$, $a_1 = \ln a$ y $b_1 = \ln b$.

La siguiente tabla contiene los valores de los datos transformados.

t	y_1	t	y_1	t	y_1
1522	7.3278	1801	7.4961	1892	7.5454
1569	7.3582	1824	7.5088	1927	7.5637
1611	7.3846	1836	7.5153	1950	7.5756
1700	7.4384	1844	7.5197	1963	7.5822
1720	7.4501	1864	7.5305	1973	7.5873
1741	7.4622	1875	7.5364	1984	7.5929
1751	7.4679	1883	7.5406	1995	7.5984
1778	7.4832	1888	7.5433	2000	7.6009

Tabla 1.4

Y el gráfico de y_1 como función de t es el siguiente.

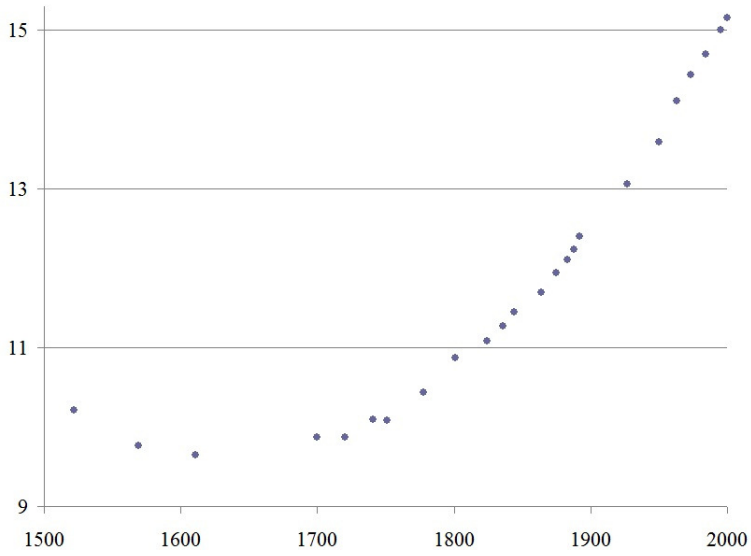

Figura 1.5 $y_1 = \ln y$ como función de t

¡Sorpresa! Tampoco la relación entre y_1 y t es lineal, de modo que la propuesta $y_1 = a_1 + b_1 t$ no es satisfactoria. El gráfico sugiere que la relación entre t y y_1 es más bien cuadrática: $y_1 = at^2 + bt + c$.

Para estimar los coeficientes a, b y c en la ecuación anterior podrían usarse técnicas de regresión múltiple, como se mencionó en la primera sección. Pero otra opción es escribir la relación cuadrática en la forma

$$y_1 = a(t-h)^2 + k$$

que es una forma alterna para la ecuación de una parábola, donde el punto (h,k) es el vértice. La ventaja de esta forma en el caso en estudio es que fácilmente se estima h de manera visual para no necesitar regresión múltiple. En efecto, se observa en el gráfico que $h \approx 1640$ (el valor de t donde se alcanza el vértice), así que la ecuación puede escribirse como

$$y_1 = ax + k$$

donde se define la nueva variable $x = (t-1640)^2$. Esta ecuación, $y_1 = ax + k$, también es lineal, pero no se puede confiar en que sea aceptable antes de ver el gráfico. Afortunadamente, en el gráfico de x vs y_1, a continuación, se nota que la relación sí es casi exactamente lineal.

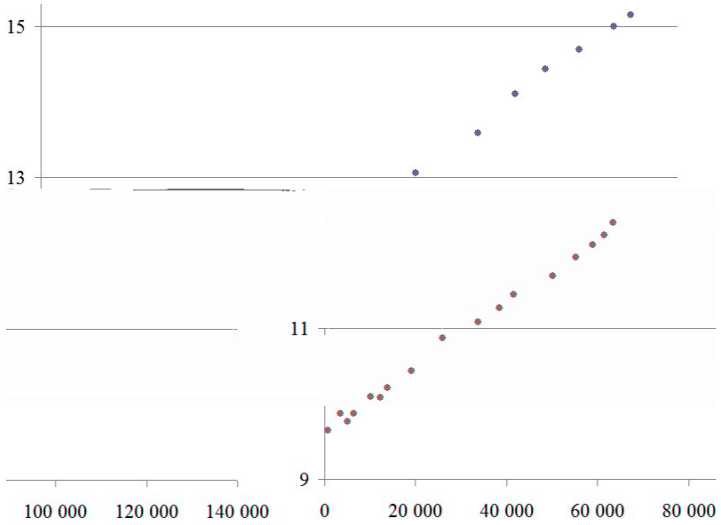

Figura 1.6 y_1 como función de $x = (t-1640)^2$

El análisis de regresión para $y_1 = ax + k$ arroja los coeficientes $a = 4.2629 \times 10^{-5}$, $k = 9.6273$. Entonces, devolviendo los cambios de variables que se hicieron, resulta

$$\begin{aligned} y_1 &= 4.2629 \times 10^{-5} x + 9.6273 \\ \ln y &= 4.2629 \times 10^{-5} (t - 1640)^2 + 9.6273 \\ y &= \exp\left[4.2629 \times 10^{-5} (t - 1640)^2 + 9.6273\right] \\ &= 15173.8 \cdot 1.00004263^{(t-1640)^2} \end{aligned}$$

(donde exp es la función exponencial natural).

El gráfico siguiente muestra los puntos que habíamos visto en la figura 1.4 junto con el gráfico de la ecuación anterior. Como se ve, la regresión es bastante precisa.

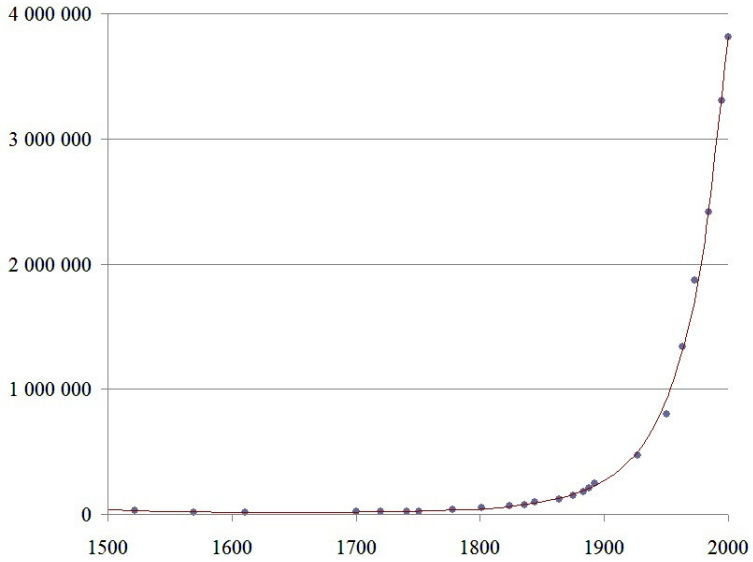

Figura 1.7 Curva de regresión para la población como función del tiempo

Finalmente, se acepta la siguiente ecuación como aproximación de la población de Costa Rica en función del año:

$$\text{Población} = 15173.8 \cdot 1.00004263^{(\text{Año}-1640)^2}$$

Bibliografía

[1] Acuña, L. (2004). *Estadística aplicada con Fathom* (1era ed). Costa Rica: Editorial Tecnológica de Costa Rica.

[2] Devore, J. (2006). *Probabilidad y estadística para ingeniería y ciencias* (6ta ed). México: Thomson Paraninfo.

Derechos Reservados © 2012 Revista digital Matemática, Educación e Internet (www.tec-digital.itcr.ac.cr/revistamatematica/)