MODELOS DE REGRESIÓN LINEALES Y NO LINEALES: SU APLICACIÓN EN PROBLEMAS DE INGENIERÍA [1]

Claudia Minnaard
Facultad de Ingeniería. Universidad Nacional de Lomas de Zamora
Universidad CAECE
Buenos Aires. Argentina.
minnaard@uolsinectis.com.ar

RESUMEN

Habitualmente el tratamiento de la regresión se limita al caso lineal. En muchos casos esto puede ser suficiente pero en otros no. Será necesario probar la linealidad de la curva de regresión, dicha prueba se puede obtener por el método de análisis de la variancia.
En el presente trabajo se describe la aplicación de modelos lineales y no lineales en problemas de ingeniería, utilizando el software XLStat. Asimismo se describe el intervalo de confianza para el coeficiente de regresión en el modelo lineal, para la ordenada al origen y para la imagen a través de la recta. En el caso de los modelos no lineales se prueba la bondad del ajuste realizado a través de las pruebas específicas.
La correcta elección de un modelo adecuado, que describa los datos en problemas de ingeniería, proporciona elementos de juicio suficientes para la toma de decisiones en condiciones de incertidumbre.

Palabras clave: regresión lineal, regresión no lineal, mejor ajuste, método de mínimos cuadrados

INTRODUCCIÓN

En el análisis de regresión[2] una de las dos variables, que llamamos **X**, puede considerarse como variable ordinaria, es decir se puede medir sin error apreciable. La otra variable **Y**, es una variable aleatoria. A **X** se la llama *variable independiente* (algunas veces variable controlada) y nuestro interés es la dependencia de **Y** en términos de **X**.

Supongamos que en cierto experimento aleatorio tratamos de manera simultánea dos variables, una variable ordinaria **X** y una variable aleatoria **Y.** Efectuamos el experimento de tal manera que

[1] Trabajo aceptado para ser presentado en forma oral en el IICaim –Segundo Congreso Argentino de Ingeniería Mecánica. San Juan, Argentina. Noviembre 2010

[2] Este término lo sugirió la observación de Galton que en promedio los hijos de padres altos no son tan altos como sus padres y los hijos de padres bajos no son tan bajos como sus padres, así que existe la tendencia a *regresar* hacia la media. El término *correlación* fue también propuesto por Galton (Proceedingd of the Royal Society of London, 45, 1888). En: http://rspl.royalsocietypublishing.org/content/by/year/1888

seleccionamos primero *n* valores $x_1, x_2, \ldots\ldots\ldots, x_n$ de **X** y luego para cada x_j obtenemos un valor observado y_j de **Y.** Entonces, tenemos una muestra de *n* parejas de valores:

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

Podemos graficar las *n* parejas como puntos del plano.

Nuestro objetivo es hallar alguna función que describa aproximadamente el diagrama de puntos anterior, en el rango considerado de la variable **X**.

A tal efecto en primer lugar elegimos una clase de funciones de donde seleccionaremos alguna función apropiada.

Las clases de funciones más utilizadas son las siguientes:

i) Polinomiales

a) Lineales

$$f(x, a) = a_0 + a_1 x \qquad a = (a_0, a_1)$$

b) Cuadráticas

$$f(a, x) = a_0 + a_1 x + a_1 x^2 \qquad a = (a_0, a_1, a_2)$$

c) En general de grado menor o igual a m

$$f(a, x) = a_0 + a_1 x + a_2 x^2 + \ldots\ldots\ldots + a_m x^m \qquad a = (a_0, a_1, a_2, \ldots\ldots\ldots, a_m)$$

ii) Potenciales

$$f(x, a) = a_0 . x^{a_1} \qquad a = (a_0, a_1)$$

iii) Exponenciales

$$f(x, a) = a_0 . (a_1)^x \qquad a = (a_0, a_1)$$

iv) Logarítmicas

$$f(x, a) = a_0 + a_1 \ln x \qquad a = (a_0, a_1)$$

Ya elegida la clase de funciones $\boldsymbol{C} = \{f(x,a)/a \in A \subseteq \Re^n\}$ nos falta determinar en la misma alguna que describa los valores dados. Para realizar tal propósito necesitamos un criterio, utilizaremos aquí el llamado ***método de mínimos cuadrados.*** Diversos autores (Sotomayor et al. (2002); Mendenhall et al. (2004), García (2004), Montgomery el al. (2002)) describen el método.

Sea $f(x,a)$ una función cualquiera de la clase $\boldsymbol{C}$. Graficando*:(Gráfico 1)*

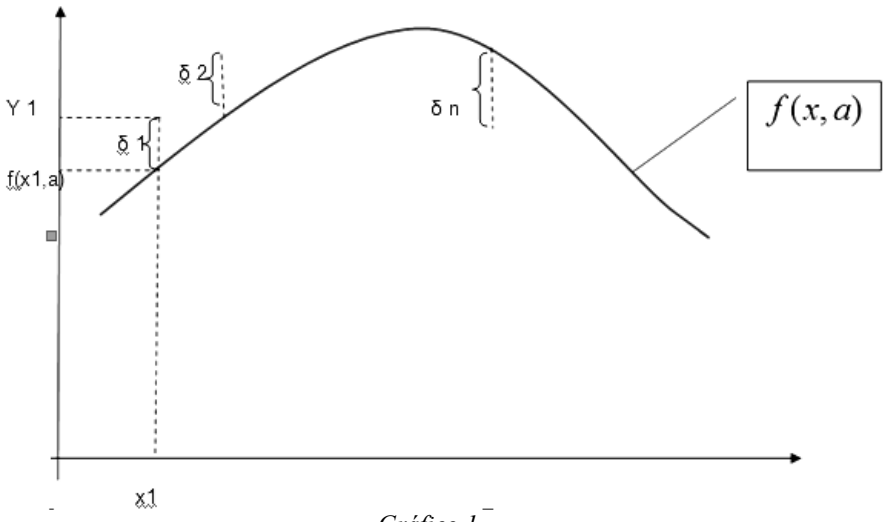

Gráfico 1

Tenemos que para cada valor de i = 1,2, ……,n el error es la diferencia entre el valor observado y el obtenido a través de la función

$$\delta_i = Y_i - f(x_i, a)$$

Definimos la función S que en cada vector vale

$$Sa = \sum_{i=1}^{n} \delta_i^{\,2} = \sum_{i=1}^{n} (Y_i - f(x_i, a))^2$$

Si los puntos $(x_1, y_1), (x_2, y_2), \ldots\ldots\ldots\ldots, (x_n, y_n)$ pertenecen a la gráfica de f entonces $Sa = 0$ y recíprocamente.

Si eso no ocurre debemos procurar que dicha función tome el valor mínimo. En general tal mínimo se alcanza para un único valor de $\hat{a}$. Entonces la función que elegimos es la función $\hat{y} = f(x, \hat{a})$ que se denomina función de mínimos cuadrados para el problema dado.

Determinación de $\hat{a}$:

Recordemos que para que *S* tenga un mínimo en el punto *a* debe ocurrir $\frac{\partial}{\partial a_j} Sa = 0$

Por lo tanto obtenemos

$$-2\sum_{i=1}^{n}(Y_i - f(x_i, a))\frac{\partial}{\partial a_j} f(x_i, a) = 0$$

De donde resulta

$$\sum_{i=1}^{n} f(x_i, a))\frac{\partial}{\partial a_j} f(x_i, a) = \sum_{i=1}^{n} Y_i \frac{\partial}{\partial a_j} f(x_i, a) \qquad \text{j =0,1,...,m}$$

Desarrollando las sumatorias resulta un sistema de ecuaciones normales cuya solución es el vector $\hat{a}$ buscado

REGRESIÓN LINEAL

Aplicar el método de mínimos cuadrados utilizando las herramientas que nos proporcionan las Tics resulta sencillo.

Tomemos un ejemplo:

En la producción de herramientas, el método para deformar acero a temperatura normal mantiene una relación inversa con la dureza del mismo ya que, a medida que la deformación crece, se ve afectada la dureza del acero. Para investigar esta relación se ha tomado la siguiente muestra:

X: deformación (en mm)	*6*	*9*	*11*	*13*	*22*	*26*	*28*	*33*	*35*
Y : dureza Brinell (en kg/mm2)	*68*	*67*	*65*	*53*	*44*	*40*	*37*	*34*	*32*

Tabla 1: Deformación y Dureza Brinell

Representamos mediante un diagrama de dispersión indicando la recta de regresión, su ecuación y el coeficiente de determinación correspondiente.[3]

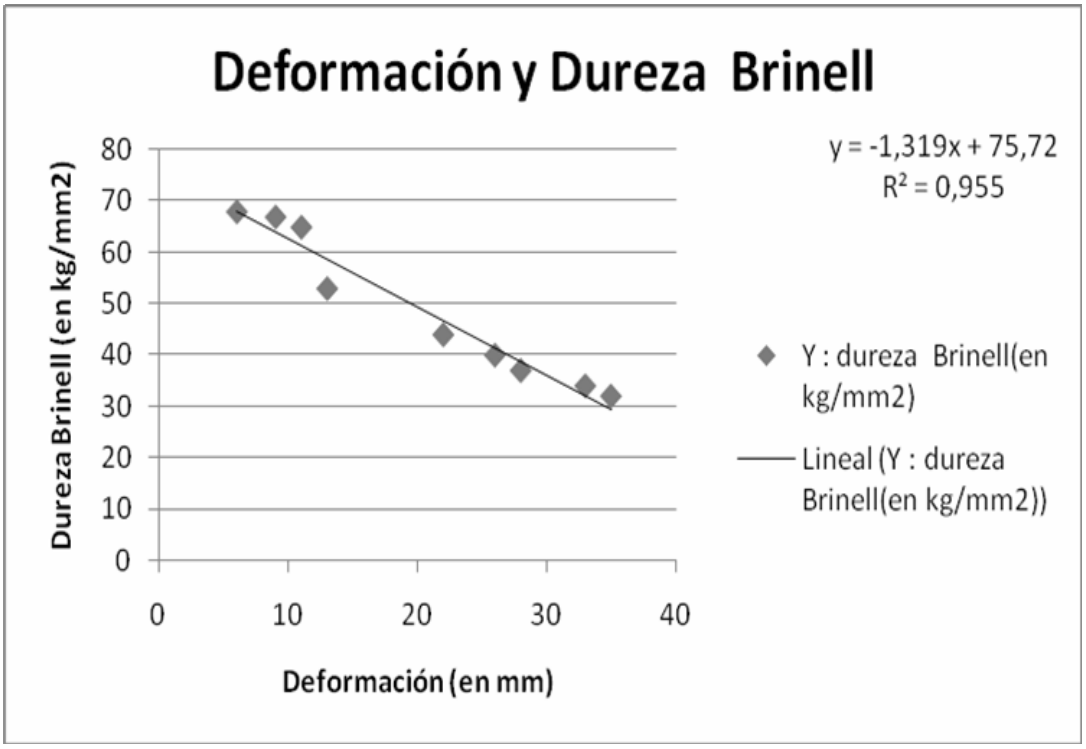

Gráfico 2: Diagrama de dispersión y recta de regresión para los datos de la Tabla 1.

ANALISIS DE LA VARIANZA

Siendo $\hat{y} = a + bx$ la recta de regresión lineal, el parámetro b se llama **coeficiente de regresión**. Su valor expresa el incremento de $\hat{y}$ cuando x aumenta una unidad. Si b toma un valor positivo, la variable $\hat{y}$ crece al crecer x , y en consecuencia la recta es creciente, lo que indica que la dependencia entre las variables es directa.

Si b toma un valor negativo, la variable $\hat{y}$ decrece al crecer x , y en consecuencia la recta es decreciente, lo que indica que la dependencia entre las variables es inversa. Si $b = 0$, la recta es horizontal y no hay dependencia entre las variables, ya que las variaciones de x no provocan variación de $\hat{y}$.

[3] Tanto el diagrama de dispersión como la recta de regresión fueron representados mediante Excel 2007.

Es muy interesante ver que los **métodos de análisis de la varianza** también pueden aplicarse en relación con problemas de regresión.

Utilizando el software XLStat[4] aplicamos el análisis de la varianza a los datos de la Tabla 1.

Prueba de la hipótesis $\beta = 0$ contra la alternativa $\beta \neq 0$ (β : coeficiente de regresión de la población)

Análisis de la varianza:

Fuente	GDL	Suma de los cuadrados	Media de los cuadrados	F	Pr > F
Modelo	1	1643,735	1643,735	149,132	< 0,0001
Error	7	77,154	11,022		
Total corregido	8	1720,889			

Calculado contra el modelo Y=Media(Y)

Tabla 2: Análisis de la varianza aplicado al modelo

Como F = 149,132 > 0,001 se rechaza la hipótesis nula, por lo tanto $\beta \neq 0$

Parámetros del modelo:

Fuente	Valor	Desviación típica	t	Pr > |t|	Límite inferior (95%)	Límite superior (95%)
Intersección	75,720	2,460	30,779	< 0,0001	69,899	81,541
Deformación	-1,320	0,108	-12,212	< 0,0001	-1,575	-1,064

Tabla 3: Parámetros del modelo lineal e intervalos de confianza del 95% para el coeficiente de regresión y la ordenada al origen

[4] Ver en http://www.xlstat.com/

EL COEFICIENTE DE DETERMINACION

Se define el **coeficiente de determinación** como la parte relativa de la variación total que viene explicada por el modelo.

$$R^2 = \frac{\overline{(\hat{y} - \bar{y})^2}}{S_y^{\ 2}}$$

- El **coeficiente de determinación** toma valores entre 0 y 1. ($0 \leq R^2 \leq 1$).
- Todo ajuste mínimo cuadrático debe venir acompañado de su respectivo **coeficiente de determinación** para poder conocer el poder representativo de la función de ajuste, es decir el valor explicativo del modelo.
- Si $R^2 > 0{,}90$ se acepta el ajuste, en caso contrario se debe buscar otro modelo.
- Para el ejemplo propuesto $R^2 = 0{,}955 > 0{,}90$ por lo tanto la regresión lineal es un muy buen ajuste.

REGRESIONES NO LINEALES

Si bien la regresión lineal es un muy buen ajuste para el problema propuesto, aplicaremos otros tipos de regresiones a fin de poder comparar.

Cuadrática

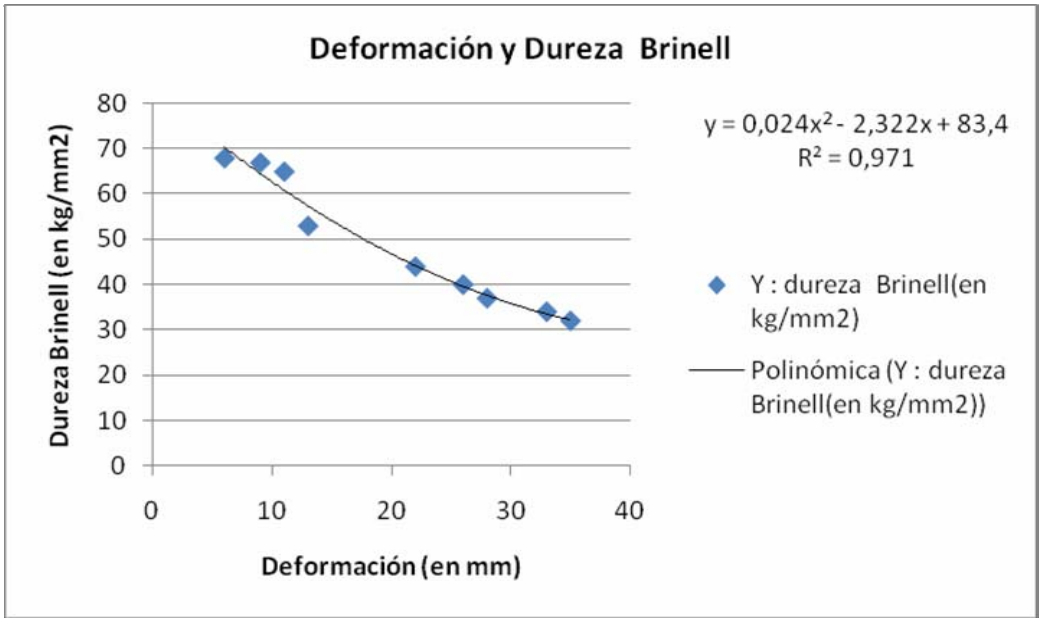

Gráfico 3 : Regresión cuadrática

Potencial

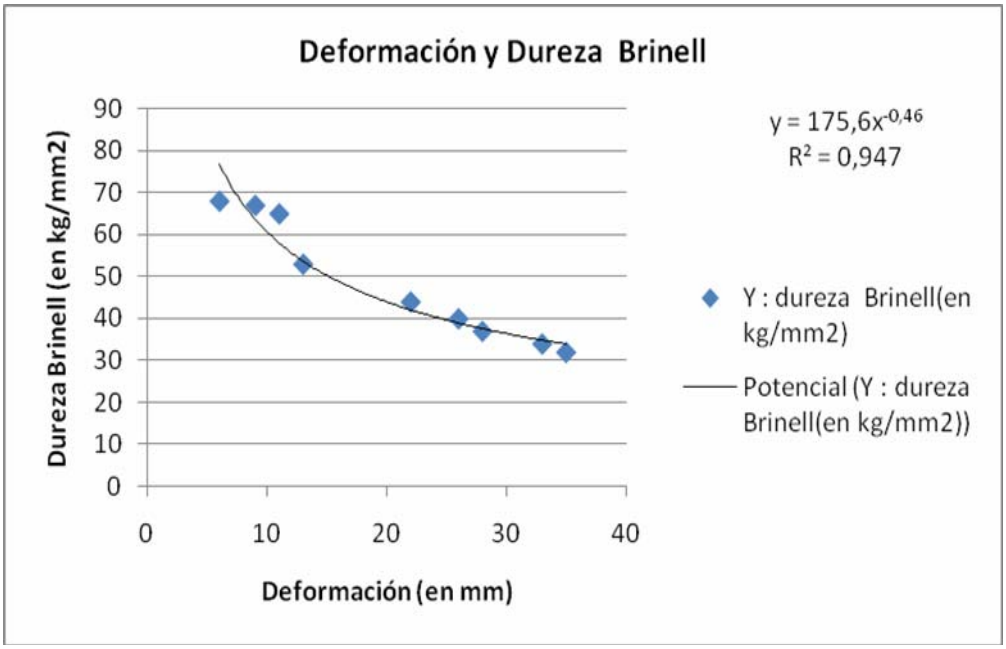

Gráfico 4 : Regresión potencial

Exponencial

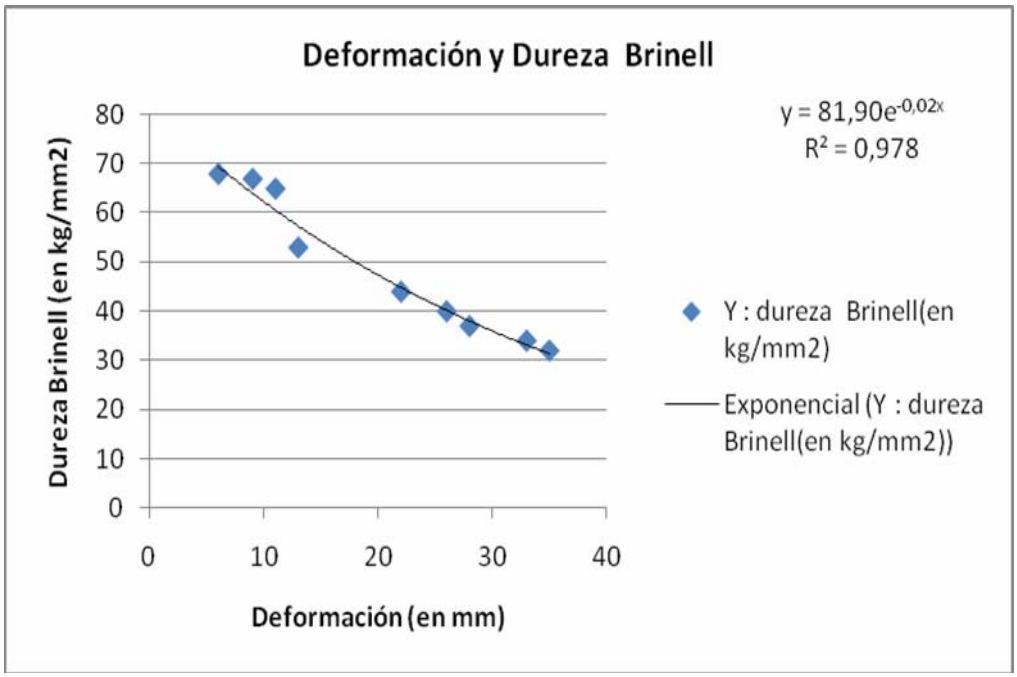

Gráfico 5 : Regresión exponecial

Logarítmica

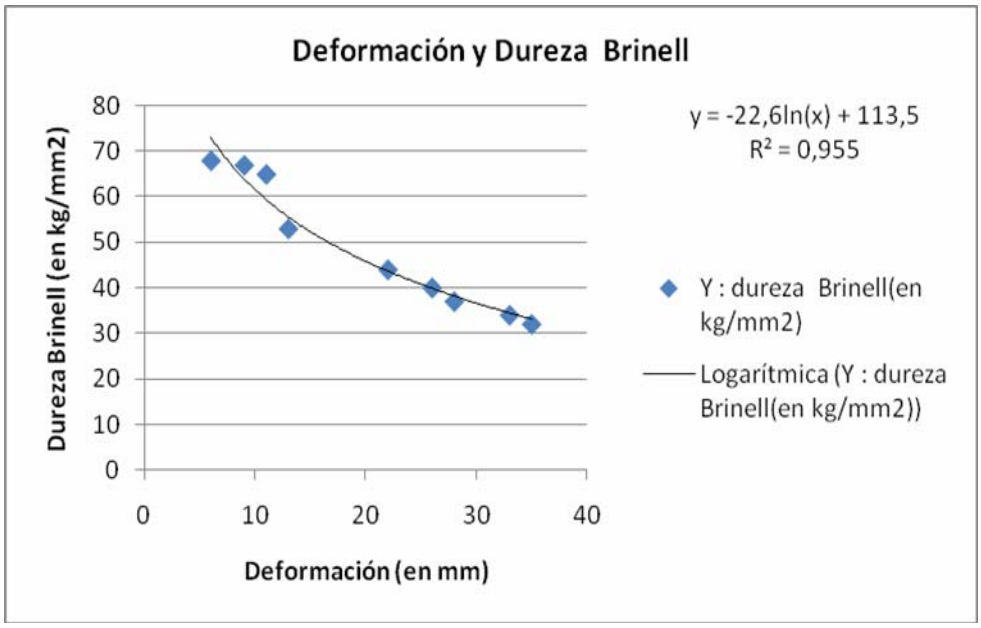

Gráfico 6 : Regresión logarítmica

Compararamos las regresiones aplicadas

Tipo de regresión	Coeficiente de determinación
Lineal	0,955
Cuadrática	0,971
Potencial	0,941
Exponencial	0,978
Logarítmica	0,955

Tabla 4: Comparación entre las regresiones

Es posible observar(*Tabla 4*) que las regresiones son muy buenas, ya que en todos los casos $R^2 > 0{,}90$.

El mejor ajuste es a través de una función exponencial ($R^2 = 0{,}978$), seguido por la regresión cuadrática ($R^2 = 0{,}971$), si bien se puede observar que los dos valores anteriores están muy cerca uno del otro.

CONCLUSIONES

La aplicación de distintas regresiones sobre un mismo problema nos permite realizar comparaciones, sin limitarse solamente al caso lineal.

La facilidad de que nos brindan las nuevas tecnologías permiten en poco tiempo efectuar comparaciones que nos permitan la correcta elección de un modelo adecuado, que describa los datos en problemas de ingeniería, así como nos proporciona elementos de juicio suficientes para la toma de decisiones en condiciones de incertidumbre.

REFERENCIAS BIBLIOGRÁFICAS

Velasco Sotomayor,G. ; Wisniewski, P.(2002) *Probabilidad y Estadística para Ingeniería y Ciencias.* Editorial Thomson.

Mendenhall,W. ; Wackerly,D.; Sheaffer, R.(2004) *Estadística Matemática con Aplicaciones.* Grupo Editorial Iberoamérica.

García, R. (2004). *Inferencia estadística y diseño de experimentos.* Buenos Aires: Eudeba.

Montgomery,D. ; Peck,E. y Vinning, G. (2002). *Introducción al Análisis de Regresión Lineal.* Editorial C.E.C.S.A.