

## Capítulo IX

# Análisis de Regresión no lineal

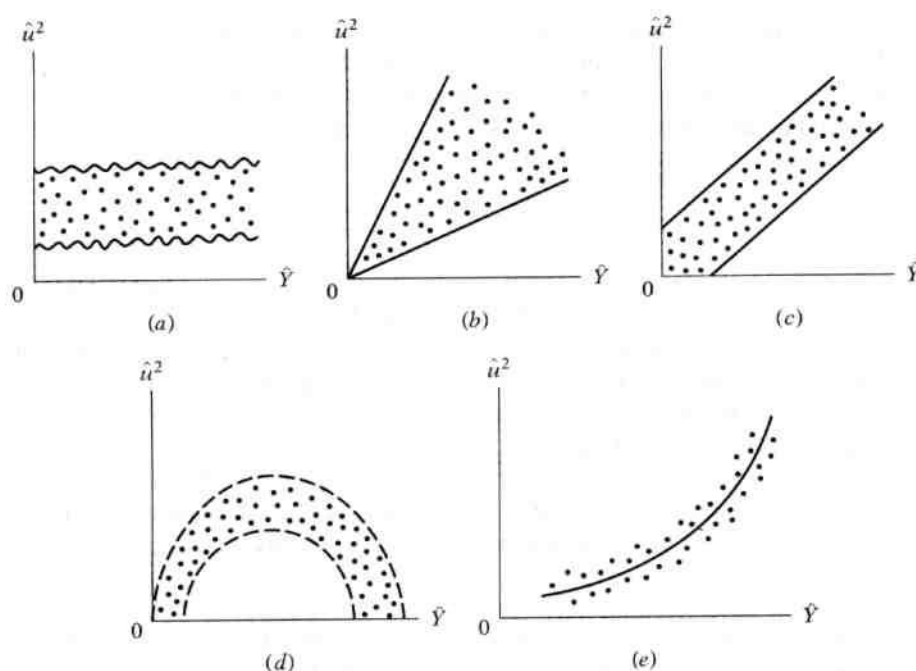
En algunas situaciones es evidente que un modelo lineal en todas las variables independientes es inadecuado. Por ejemplo, un modelo de regresión que predice  $Y$ , la puntuación de las preferencias en una prueba de sabor de una bebida de lima, como función lineal de  $X_1$  (concentración de lima) y  $X_2$  (dulzura) es muy dudosa. En otros casos, los diagramas de dispersión de los datos revelan la ausencia de linealidad. Particularmente, en la regresión lineal, una gráfica ordinaria de  $Y$  contra  $X$  es adecuada.

En la regresión múltiple, el efecto de las variables entre sí puede oscurecer la ausencia de linealidad. Por esta razón, una estrategia estándar es ajustar un modelo de primer orden y después representar gráficamente los residuos de este modelo con cada variable independiente. Si un modelo más apropiado contiene, por ejemplo, un término en segundo grado  $X_1^2$ , entonces la gráfica de los residuos ( $\varepsilon = Y_i - \hat{Y}$ ) contra  $X_1$  muestra un patrón no lineal. Como el uso de un modelo de primer orden elimina el efecto lineal de las otras variables independientes, las no lineales a menudo se muestran con mayor claridad en estos diagramas residuales.

Las interacciones entre las variables son más difíciles de descubrir en los diagramas de dispersión. Si  $X_1$  y  $X_2$  interactúan al determinar  $Y$ , hay tres variables involucradas; desafortunadamente, los diagramas tridimensionales son más difíciles de trazar e interpretar. Quizá el sentido común sea la mejor manera de determinar si hay interacciones presentes.

Para el caso especial en que una de las variables independientes es una variable cualitativa representada por una o más variables ficticias (dummy), la interacción se puede descubrir trazando gráficas de los residuos (tomados de un modelo de primer orden) contra otras variables independientes. Se deberían trazar gráficas por separado para las observaciones en cada categoría de la variable cualitativa. El modelo de primer orden sin interacciones implica que estos diagramas deben ser paralelos. Si los diagramas de los residuos por separado no son más o menos paralelos, se deberían considerar la posible presencia de algún tipo de interacción.

La representación gráfica de estos residuales con respecto al valor estimado de  $Y$ , pueden producir los siguientes patrones:

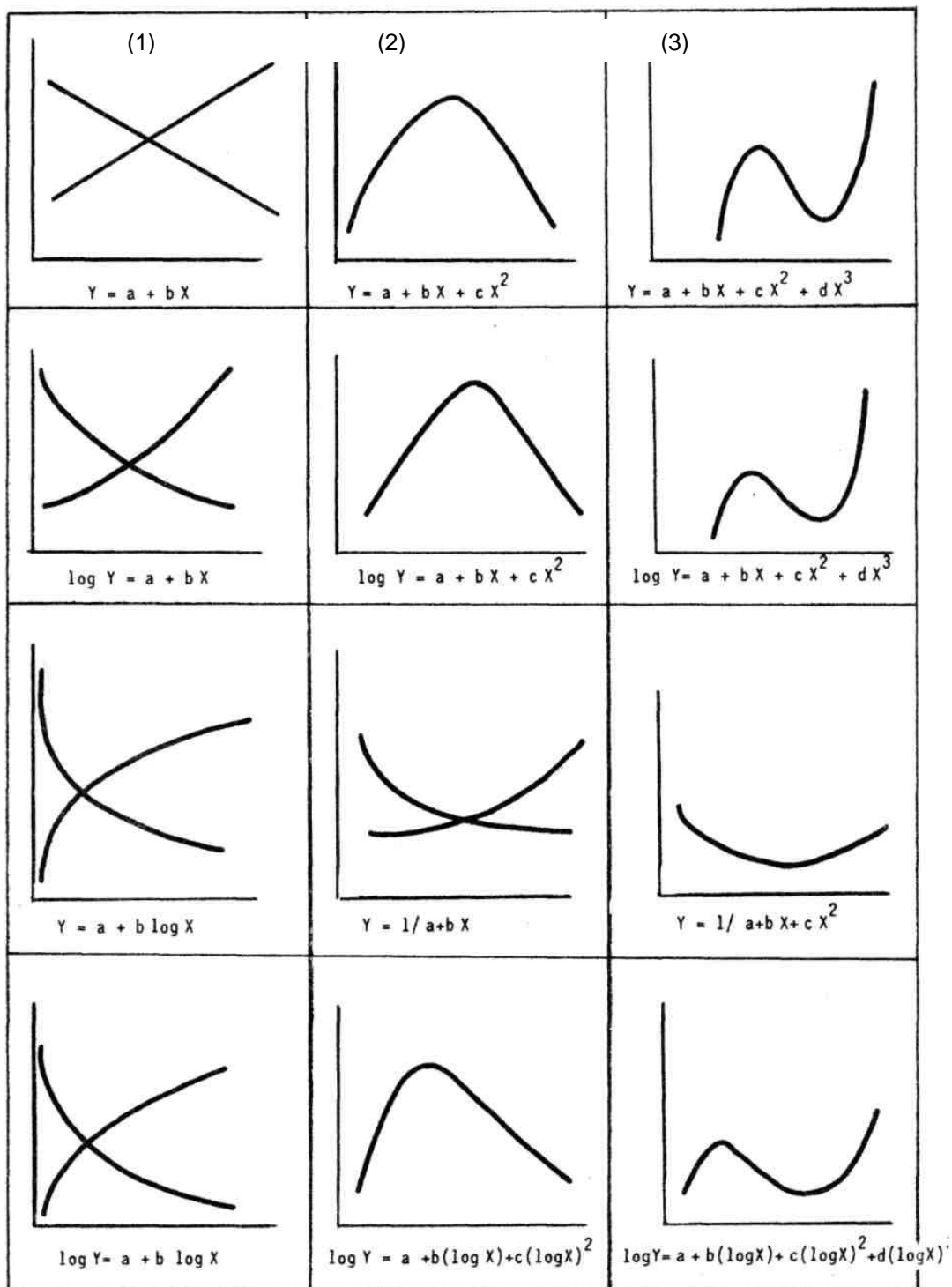


La idea de estos gráficos es averiguar si el valor medio estimado de  $Y$  está relacionado sistemáticamente con el residual al cuadrado.

En (a) se ve que no hay un patrón sistemático entre las dos variables, lo cual sugiere que posiblemente no hay heteroscedasticidad en los datos. Sin embargo las gráficas de (b) a (e) muestran patrones

definidos. Por ejemplo, la figura (c) sugiere una relación lineal, mientras que (d) y (e) indican una relación cuadrática. Utilizando estas gráficas, es posible obtener curvas de mejor ajuste u orientarnos acerca del tipo de transformación que debe aplicarse a los datos a fin de mejorar el ajuste. Pueden también hacerse representaciones de los residuales versus el valor de  $X$ .

Los modelos de regresión no lineal son muy variados. Algunos de estos modelos se dan a continuación:



En la columna (1) aparecen de arriba abajo el modelo lineal, el semilogarítmico inverso en Y, el modelo inverso logarítmico en X, doble logarítmico. En la columna (2) aparecen el modelo cuadrático, el modelo cuadrático inverso logarítmico en Y, el modelo recíproco, el modelo de parábola logarítmica y en la tercera columna aparecen modelos polinomiales.



Existen una gama muy amplia de modelos. Pero se pueden clasificar en dos grupos los intrínsecamente lineales, es decir, se pueden transformar a la forma lineal y los no lineales propiamente dichos que no se pueden transformar a la forma lineal. Los métodos de estimación cambian en ambos casos en los intrínsecamente lineales puede aplicarse la transformación y luego aplicar el método de los mínimos cuadrado. En los no lineales propiamente dichos se aplica el método de estimación de máxima verosimilitud.



Entre los modelos intrínsecamente lineales más usuales en la práctica tenemos:

- (a) El modelo lineal** útil en los casos de ajuste no lineal porque sirve como patrón de comparación.

Ecuación del modelo  $\hat{Y} = a + bX$

- (b) El modelo recíproco** (también llamado hipérbola) donde una de las variable va aumentado y la otra va disminuyendo.

Ecuación del modelo  $\hat{Y} = \frac{1}{a + bX}$

**(c) El modelo gamma** que es muy utilizado en ajustes de curvas de producción lechera.

Ecuación del modelo  $\hat{Y} = ab^X X^c$  su transformación a la forma lineal es:

$$\log Y = \log(a) + x(\log b) + c(\log X)$$

**(d) El modelo potencial** muy utilizado en ajusten de precio-demanda.

Ecuación del modelo  $\hat{Y} = aX^b$  su transformación a la forma lineal es:

$$\log Y = \log(a) + b(\log X)$$

**(e) El modelo exponencial** muy utilizado en ajustes de crecimiento poblacionales.

Ecuación del modelo  $\hat{Y} = ab^X$  su transformación a la forma lineal es:

$$\log Y = \log(a) + X(\log b)$$

Entre los modelos no lineales propiamente dichos encontramos:

**(a) El modelo logístico** para estudiar el crecimiento de poblaciones.

Ecuación del modelo  $\hat{Y} = \frac{1}{k + ab^X}$  o  $\frac{1}{Y} = ab^X + k$

**(b) El modelo de parábola logarítmica**, cuya ecuación del modelo es:

$$\hat{Y} = a + b(\log X) + c(\log X)^2$$

**(c) El modelo de Gompertz** también usado para el estudio de crecimientos poblacionales.

Ecuación del modelo:  $\hat{Y} = pq^{b^x}$  o  $\log Y = \log(p) + b^x \log(q)$

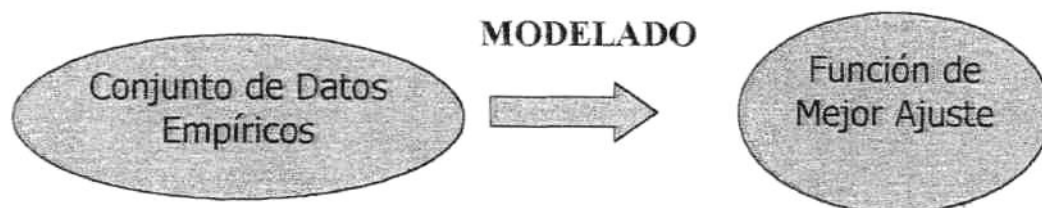
### **Modelaje, Modelado o Modelización**

Es la construcción de un modelo que se sabe a ciencia cierta, la respuesta que produce, luego se aplica a una situación donde dicha respuesta es desconocida para conocer su comportamiento, y, si el modelo resulta ser de "buen ajuste en lo sucesivo lo aplicaremos para propósitos de predicción.

**Modelado Matemático-Estadístico:** es una función que describe un proceso o fenómeno.

El tipo o función queda determinado por:(a) Gráfica de la función;

(b) Algún Principio subyacente a los datos(c)Mejor Criterio de ajuste.

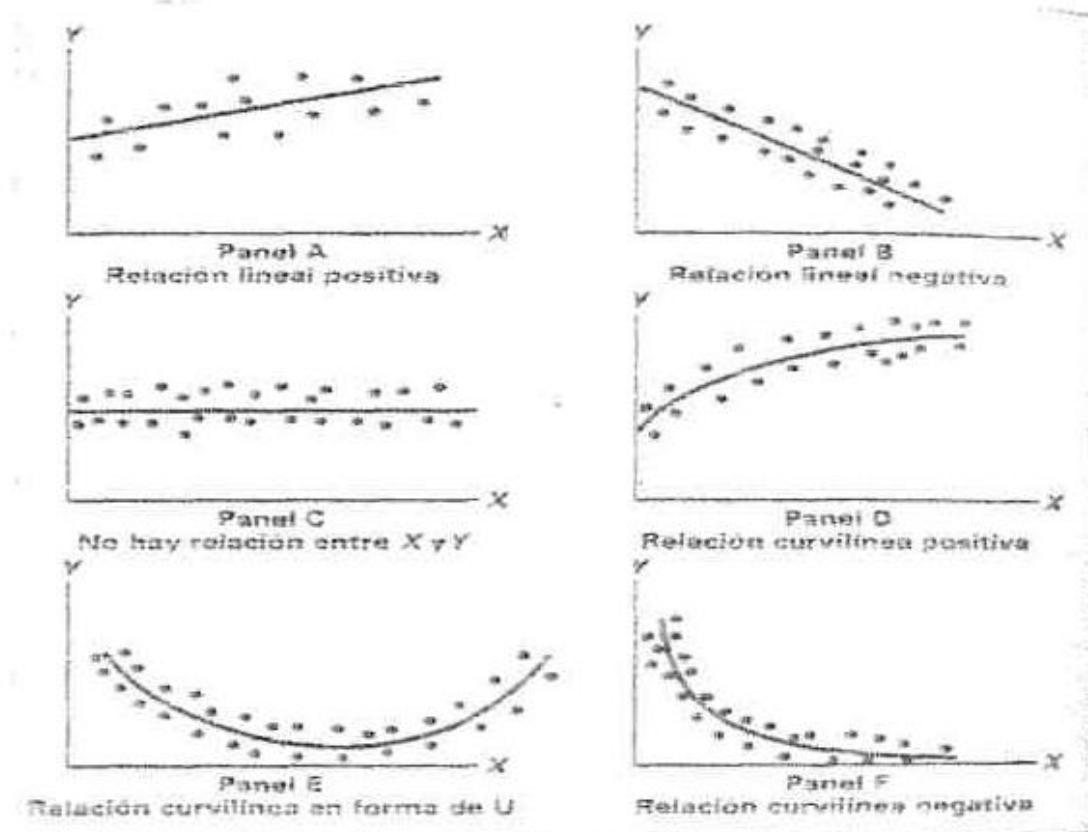


Buen ajuste

Es el grado de semejanza entre las respuestas reales ( $Y_i$ ) y las obtenidas con el modelo

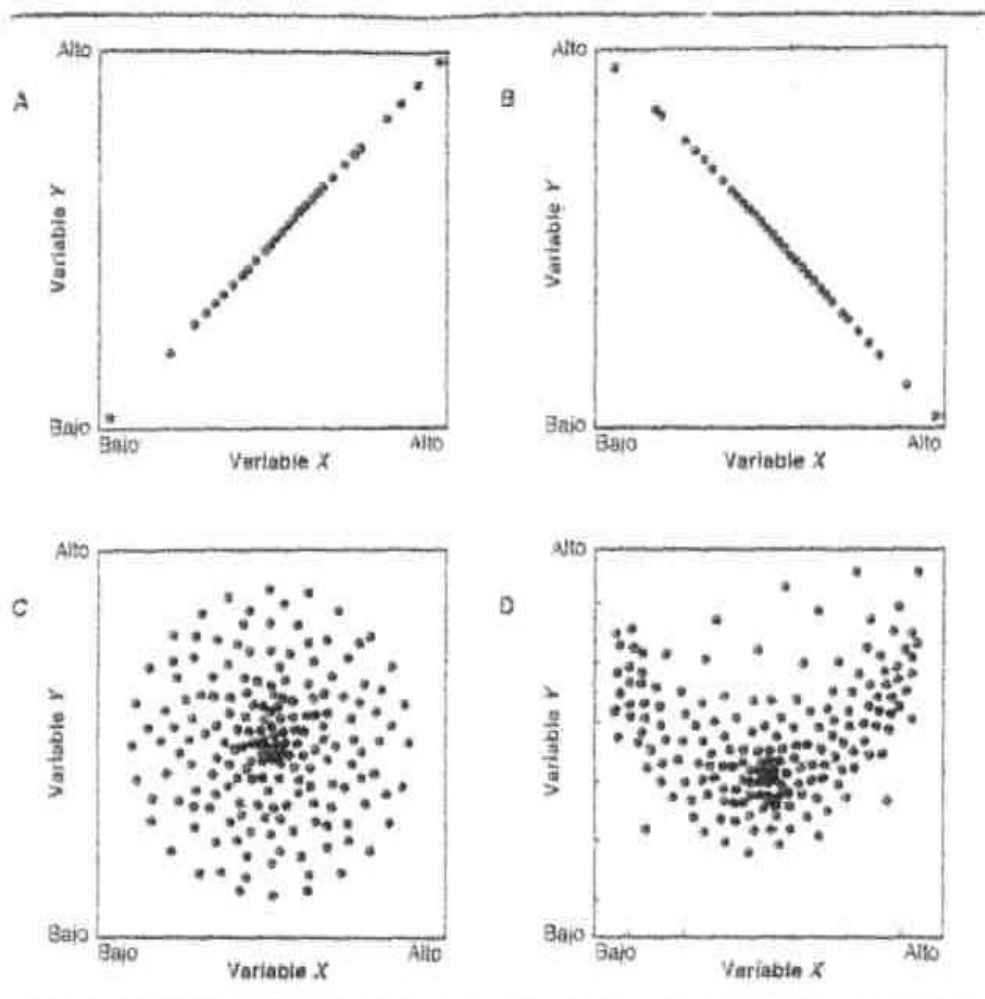
### Tipos de pendientes

## Tipos de Pendientes



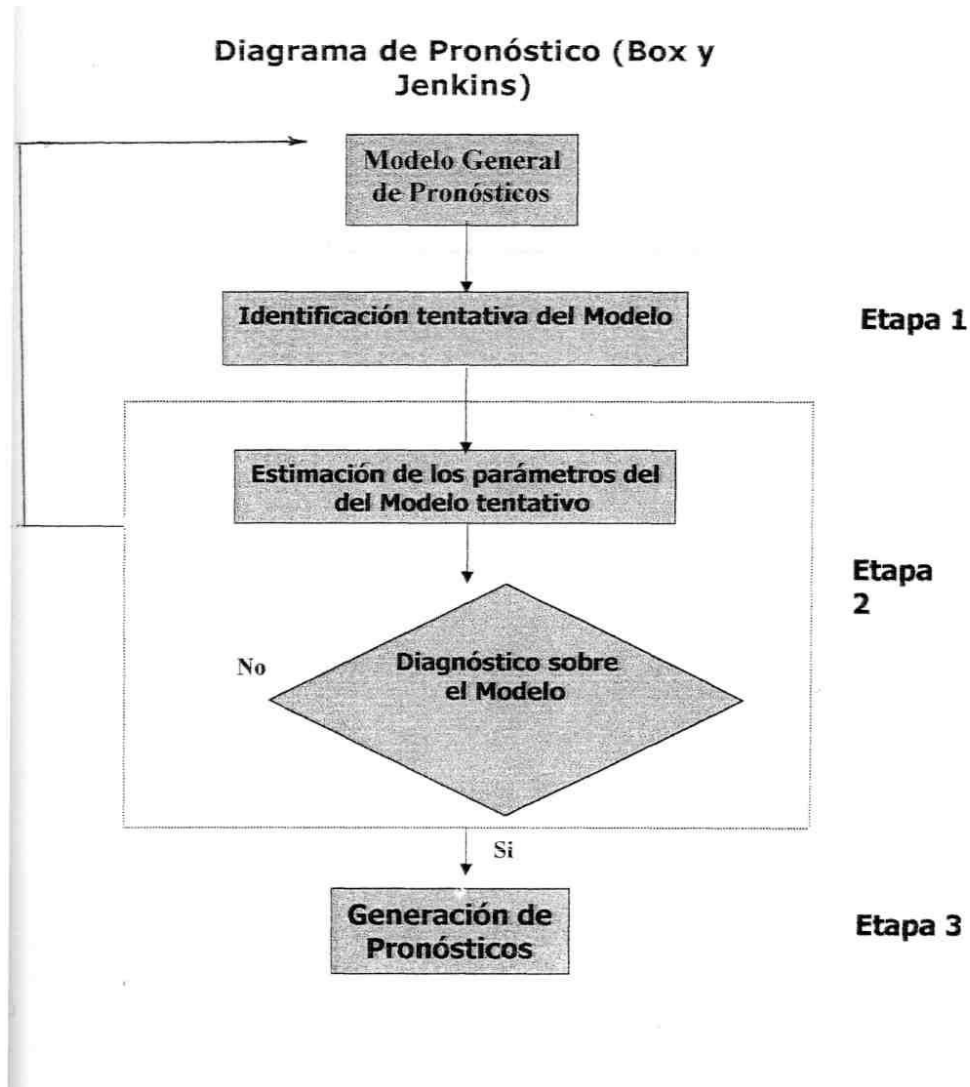
## Tipos de Diagramas de Dispersión





### Definición de Mejor Modelo:

- Más fiable (o confiable)
- Menor error estándar en los estimadores
- Más sencillo (Principio de parsimonia): Menos parámetro
- Menos suposiciones para construirlo





### **Pautas para usar la ecuación de regresión lineal**

- (1) Si no hay una correlación lineal significativa, no use la ecuación de regresión para hacer predicciones.
  - (2) No use la ecuación de regresión para hacer predicciones fuera del rango de valores muestreados.
  - (3) Una ecuación de regresión basada en datos viejos (obsoletos) no necesariamente sigue siendo válida en el presente.
  - (4) No haga predicciones acerca de una población distintas de la población de la cual se extrajo la muestra de datos.
-

## Ejemplo de Aplicación

Mediante un ejemplo relacionado con el crecimiento (longitud:Y) de una larva de un parásito por días de eclosión (X) se ajustarán una serie de modelos y se escogerá el que sea de “mejor ajuste” en base al criterio del R-cuadrado.

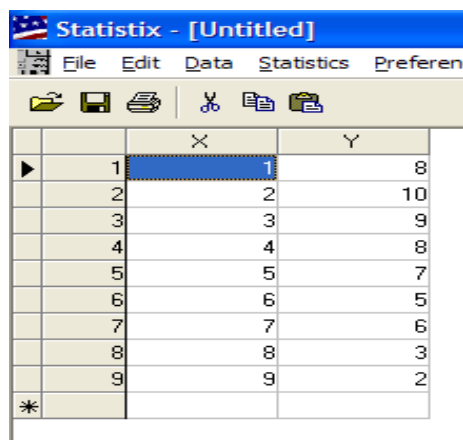
X (días)	1	2	3	4	5	6	7	8	9
Y (longitud)	8	10	9	8	7	5	6	3	2

Se pide ajustar los modelos siguientes:

- (a) Lineal
- (b) Exponencial
- (c) Potencial
- (d) Recíproco
- (e) Inverso
- (f) Gamma
- (g) Polinomial trigonométrico
- (h) Parábola
- (i) Parábola de raíz
- (j) Logarítmico en X
- (k) Logarítmico en Y
- (l) Cúbico

Solución

- (a) Ajuste del modelo lineal: comenzamos por vaciar los datos en el paquete STATISTIX.



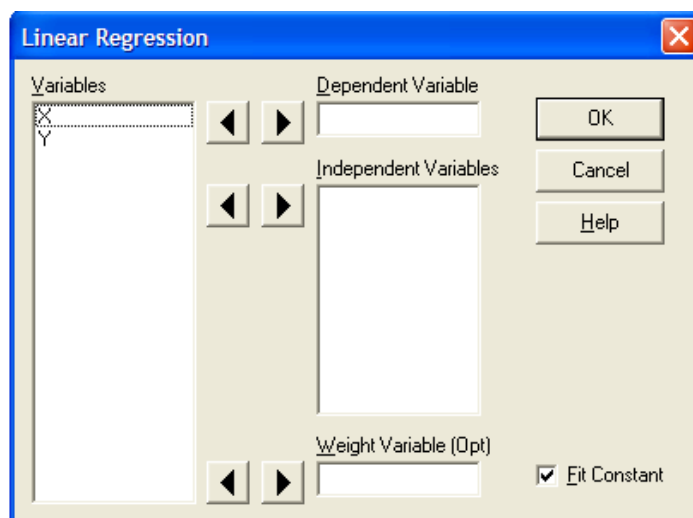
The screenshot shows the Statistix software window titled "Statistix - [Untitled]". It features a menu bar with "File", "Edit", "Data", "Statistics", and "Preferen". Below the menu is a toolbar with icons for opening, saving, printing, and other functions. The main area displays a data table with two columns, X and Y, and 10 rows of data. The first row is highlighted in blue.

	X	Y
1	1	8
2	2	10
3	3	9
4	4	8
5	5	7
6	6	5
7	7	6
8	8	3
9	9	2
*		

Luego se sigue la secuencia de instrucciones

Statistix >linear models>linear regression

Seguidamente aparece el siguiente menu de dialogo.



Declaramos Y como variable dependiente y X como la independiente.

Al presionar okay se produce la salida siguiente.

Statistix - [Linear Regression - Coefficient Table]					
File Edit Results Window Help					
Statistix 8.0 07/08/2007, 07:29:10 a.m.					
Unweighted Least Squares Linear Regression of Y					
Predictor Variables	Coefficient	Std Error	T	P	
Constant	10.9444	0.85175	12.85	0.0000	
X	-0.90000	0.15136	-5.95	0.0006	
R-Squared	0.8347	Resid. Mean Square (MSE)		1.37460	
Adjusted R-Squared	0.8111	Standard Deviation		1.17243	
Source	DF	SS	MS	F	P
Regression	1	48.6000	48.6000	35.36	0.0006
Residual	7	9.6222	1.3746		
Total	8	58.2222			
Cases Included 9 Missing Cases 0					

### El modelo lineal es:

$$Y = 10,9444 - 0,90000X \text{ con un } R^2 = 0,8347$$

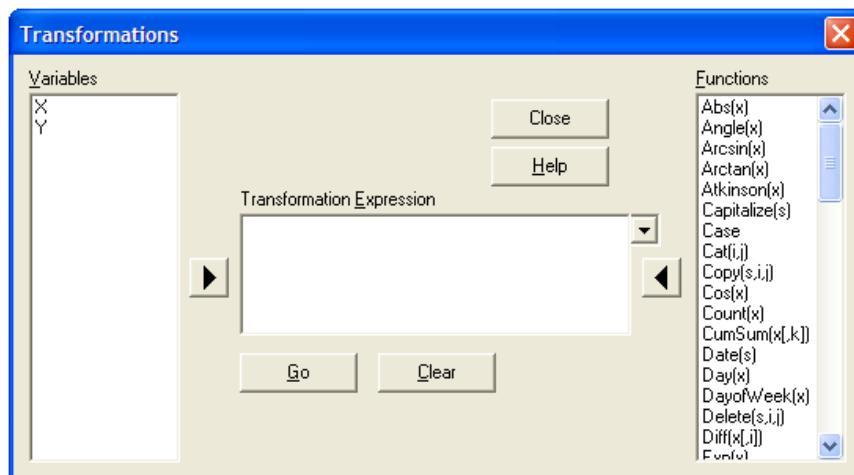
Podemos reescribir la ecuación así:

$$\text{Longitud} = 10,9444 - 0,9 \text{ Días con un } R^2 = 0,8347$$

(b) Ajuste del modelo exponencial. Como este modelo tiene la siguiente expresión:

$$Y = ab^x \text{ para expresarlo en la forma lineal es:}$$

$\log Y = \log(a) + X(\log b)$ , es decir, para introducir la data en STATISTIX debemos crear la variable LOG(Y). En este momento acudimos a la opción TRANSFORMATION del menú DATA y aparece la siguiente caja de dialogo:



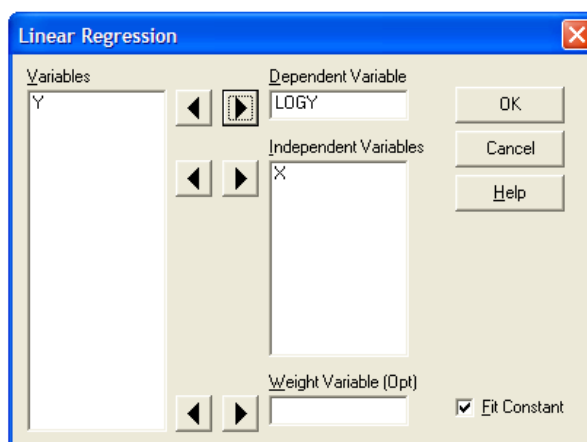
Y en el interior de la caja que dice Transformation Expresión  
Escribimos:

$\text{LOGY} = \log(Y)$

Es decir, estamos obteniendo el logaritmo de Y y se lo estamos asignando a una variable llamada LOGY. Procedemos a presionar GO y se crea esta variable; a continuación volvemos a invocar el procedimiento para la regresión:

STATISTICS>LINEAR MODELS>LINEAR REGRESSION

A continuación aparece y pasamos con el cursor LOGY como variable dependiente y X como independiente así:



Al presionar OKEY aparece la salida:

Statistix - [Linear Regression - Coefficient Table]					
File Edit Results Window Help					
Statistix 8.0 07/08/2007, 07:54:56 a.m.					
Unweighted Least Squares Linear Regression of LOGY					
Predictor Variables	Coefficient	Std Error	T	P	
Constant	1.14008	0.08495	13.42	0.0000	
X	-0.07555	0.01510	-5.01	0.0016	
R-Squared	0.7816	Resid. Mean Square (MSE)		0.01367	
Adjusted R-Squared	0.7504	Standard Deviation		0.11693	
Source	DF	SS	MS	F	P
Regression	1	0.34250	0.34250	25.05	0.0016
Residual	7	0.09571	0.01367		
Total	8	0.43820			
Cases Included 9 Missing Cases 0					

Luego la ecuación del modelo es:

$$\text{LOGY} = 1,14008 - 0,07555 X \text{ con un } R^2 = 0,7816$$

Para no hacer tan extenso el procedimiento, por demás repetitivo, para el modelo potencial (se crea la variable LOGX y LOGY), para desarrollar el modelo recíproco se crea la variable inversa de Y (invY=1/y), para desarrollar el modelo inverso creamos la variable inverso de X (invX=1/X), para desarrollar el modelo gamma se crean las variables: LOGY y LOGX, para el modelo polinomial trigonométrico se crean dos variables SENOX y COSX, así:

SENOX=sin(2\*PI\*X/24) y COSX=cos(2\*PI\*X/24), para la parábola se crea la variable X cuadrado X2=X\*X, para desarrollar el modelo parábola de raíz se crea la variable RAIZX=SQRT(X), para desarrollar el modelo cúbico se crea la variable X cubo, así X3=X^3 ó X3=X\*X\*X.

El desarrollo de estos modelos de regresión, aparecen resumido en el cuadro que sigue mostrando cada modelo y su respectivo R-cuadrado:



Modelo	R-Cuadrado
1.Lineal	0,8608
2.Cuadrático	0,8632
3.Logarítmico en X	0,6544
4.Cúbico	0,9643
5.Inverso	0,2002
6.Potencial	0,8560
7. Exponencial	0,8590
8.Recíproco	0,8613
9.Gamma	0,9256
10.Polinomial trigonométrico	0,9238

Se aprecia que entre todos los modelos el de mejor ajuste es el modelo de regresión cúbico por poseer el mayor R-cuadrado.

Restaría verificar si el modelo cúbico es bueno para predecir o estimar comparado los valores de Y observados con los de Y esperados y esto puede hacerse utilizando el Ji-cuadrado. En este sentido, se plantea el sistema hipotético:

Ho: Los valores de Y observados concuerdan con los esperados, y por tanto, el modelo es bueno para predecir.

H1: Los valores de Y observados no concuerdan con los esperados, y por tanto, el modelo es malo para predecir.

