4. Análisis de Regresión Curvilínea

Los siguientes análisis muestran como resolver un conjunto de situaciones de regresión no lineal entre variables reduciéndolas a situaciones de regresión lineal mediante el uso de transformaciones de la variable independiente, de la dependiente o de ambas.
Para facilitar la determinación del ajuste correcto en los ejemplos siguientes se ha construido artificialmente una variable dependiente como función no lineal con coeficientes conocidos de una determinada variable independiente. Esto es solo un recurso didáctico que garantiza que la variable dependiente describe exactamente el tipo de función que se aplica en cada análisis. En un análisis real la variable dependiente viene dada (y no es una función creada arbitrariamente a partir de los datos de la variable independiente). Además, en una situación real obviamente el primer problema consiste en determinar cual es la función adecuada, lo que puede implicar en un primer diagnóstico representar gráficamente la relación entre las variables mediante diagramas de dispersión y ensayar diversos modelos de función de regresión no-lineal y lineal entre los que parezcan indicados, para optar por el más adecuado a los datos teniendo en cuenta la teoría conocida sobre la relación entre esas variables.
Para estos cálculos es necesario un programa que trabaje en doble precisión en los cálculos críticos. En BMDP esto puede hacerse utilizando el 9R (evitando que efectúe una búsqueda de mejores predictores). Otros programas como el 1R o el 2R pueden producir resultados más imprecisos en estas circunstancias y no son recomendables para este tipo de análisis. El 9R tiene la ventaja adicional de permitir plots determinados por el usuario que permiten obtener diagramas de dispersión entre las variables, incluyendo residuales, valores pronosticados y otros estadísticos relativos a los residuales.

1. Función cuadrática.

En los siguientes datos disponemos de una variable independiente X y de una variable dependiente Y1 (creada por transformación como función cuadrática de X para garantizar el ajuste mediante una función cuadrática). Vamos a ensayar la estimación de la función cuadrática reduciéndola a una función lineal.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 Y1 = 13.5 + 14*X - 0.5*X2.
/REGRESSION
 DEPENDENT = Y1.

INDEPENDENT = X, X2.


 METHOD = NONE.


/PRINT DATA.

/PLOT
 XVAR = X.
 YVAR = Y1.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

Crea una transformada X2 igual a X al cuadrado.

Crea una Y1 función cuadrática de X con coeficientes conocidos fijados arbitrariamente, como recurso didáctico para garantizar el ajuste de una ecuación

Una vez creada X2 esta se introduce en la ecuación como independiente, junto a X para permitir a 9R estimar los coeficientes de una función cuadrática como si fuera lineal.

METHOD en 9R fija el método por el que se seleccionan subconjuntos de predictores. Al fijar METHOD=NONE no se utiliza ningún método de elección de soluciones de predictores y el programa actúa simplemente calculando la ecuación solicitada

El programa 9R nos permite obtener un listado de los datos de las variable utilizadas.

Una prestación muy interesante del 9R: Nos permite obtener diagramas de dispersión entre cualesquiera pares de variables que se especifiquen (uno o más pares) incluidos residuales, valores pronosticados y otros

OUTPUT SELECCIONADO

```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

DATA AFTER TRANSFORMATIONS
  NO. X        X2       Y1
----- -------- -------- --------
    1    10.00   100.00   103.50
    2     3.00     9.00    51.00
    3     2.00     4.00    39.50
    4     4.00    16.00    61.50
    5     5.00    25.00    71.00
    6     7.00    49.00    87.00
    7     6.10    37.21    80.29
    8     0.01  1000E-7    13.64
    9     0.20     0.04    16.28
   10     1.50     2.25    33.38
   11    10.00   100.00   103.50
   12     0.50     0.25    20.37
   13     0.70     0.49    23.05
   14  .004000  1600E-8    13.56
   15     3.50    12.25    56.38 (El listado sigue en la página siguiente).
   16     5.60    31.36    76.22
```

En ocasiones expresa el dato en notación científica, por razones de espacio en el formato de presentación.

```
   17     7.60     57.76     91.02
   18     8.90     79.21     98.49
   19     2.90      8.41     49.90
   20  2000E-7  400E-10     13.50

NUMBER OF CASES READ. . . . . .      20


SUMMARY STATISTICS FOR EACH VARIABLE
------------------------------------
                              STANDARD     COEFFICIENT    SMALLEST       LARGEST
  VARIABLE      MEAN          DEVIATION    OF VARIATION    VALUE          VALUE
------------ ------------ ------------ ------------- ------------ -----------

  1 X             3.92571      3.43363       0.874651      0.00020     10.00000
  2 X2           26.61151     33.57645       1.261727      0.00000    100.00000
  3 Y1           55.15419     32.38452       0.587163     13.50280    103.50000
```

Estadísticos descriptivos de las variables y matriz de correlaciones.

```
CORRELATIONS
------------
              X         X2        Y1
                    1         2         3

X           1     1.000
X2          2     0.957     1.000
Y1          3     0.989     0.901     1.000


FIRST DIGITS OF CORRELATIONS
----------------------------

   1 X         *
   3 Y1        9*
   2 X2        99*

STATISTICS FOR 'BEST' SUBSET
----------------------------
SQUARED MULTIPLE CORRELATION    1.00000
MULTIPLE CORRELATION            1.00000
ADJUSTED SQUARED MULT. CORR.    1.00000
RESIDUAL MEAN SQUARE           0.000000
STANDARD ERROR OF EST.         0.000001
F-STATISTIC                  **********
NUMERATOR DEGREES OF FREEDOM          2
DENOMINATOR DEGREES OF FREEDOM       17
SIGNIFICANCE (TAIL PROB.)        0.0000
```

Dado que la variable dependiente ha sido "creada " como función cuadrática de X mediante transformaciones, obviamente su ajuste como función cuadrática de X es perfecto.
La ecuación con sus coeficientes calculados es:

$$Y1' = 13{,}5 + 14X - 0{,}5X^2$$

La tolerancia de una variable es la proporción de su varianza NO explicada por las otras variables independientes de la ecuación. Una tolerancia muy baja indica poca contribución original al modelo. En este tipo de ecuaciones es normal encontrar tolerancias muy bajas pues cualquier variable presenta una alta correlación con

```
                                                                    CONTRI-
  VARIABLE      REGRESSION       STANDARD  STAND.     T-   2TAIL     TOL-   BUTION
NO.   NAME     COEFFICIENT          ERROR   COEF.   STAT.   SIG.   ERANCE  TO R-SQ

   INTERCEPT       13.5000 0.59690889E-6   0.417 2.3E+7   0.000
  1 X              14.0000 0.32701899E-6   1.484 4.3E+7   0.000 0.085046  0.19151
  2 X2            -0.500000 0.33441928E-7 -0.518-1.5E+7   0.000 0.085046  0.02336

SUMMARY STATISTICS FOR RESIDUALS
--------------------------------

(CASES WITH POSITIVE WEIGHT)
AVERAGE RESIDUAL                          0.0000
RESIDUAL MEAN SQUARE                 0.00000000
AVERAGE DELETED RESIDUAL                  0.0000
AVE. SQUARED DELETED RESIDUAL
   (PREDICTION MEAN SQUARE)          0.00000000
SERIAL CORRELATION                       -0.2559
DURBIN-WATSON STATISTIC                   2.5034
```

Uno de los supuestos del análisis de regresión lineal es que los residuales de las observaciones no están correlacionados serialmente (cada uno con el siguiente). El estadístico de Durbin-Watson es un test de la autocorrelación (correlación serial) de los residuales. Si los residuales no están correlacionados este estadístico vale 2. Si los residuales autocorrelacionan positivamente el estadístico se aproxima a 0. Si los residuales presentan una correlación serial negativa el estadístico se acerca a 4.
La autocorrelación es la correlación de la serie de residuales consigo misma retrasada cierto número de unidades (p.e. retrasada un caso: correlación de cada caso con el siguiente).
Aquí el valor de 2.5 indica una cierta correlación negativa, que en concreto es de -0.2559.
Estos estadísticos se aplican con el mismo

El programa ofrece además, una serie de diagnósticos, particularmente de los residuales.

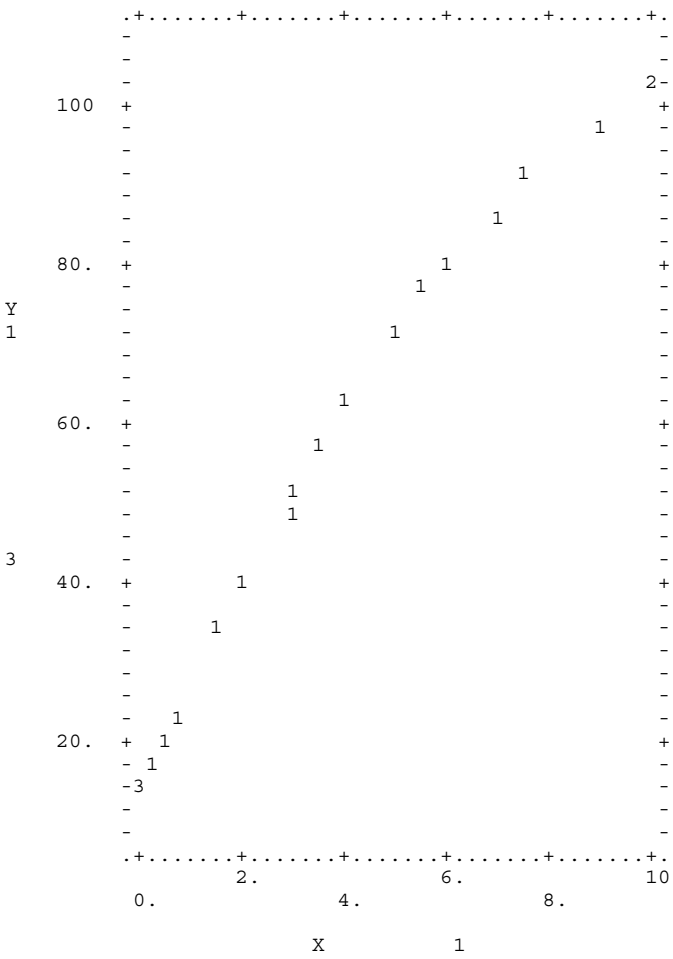


Diagrama de dispersión entre X e Y1. Obsérvese que la relación para este rango de X es casi lineal, solo una ligera deceleración en el extremo superior sugiere una curvatura característica de una relación no-lineal. Con un rango de X más amplio el efecto de curvatura se hubiera puesto de manifiesto más claramente. Para esta zona de X (entre 0 y 10 aproximadamente) probablemente una inspección visual del diagrama de dispersión no hubiera sido suficiente para pensar en una función cuadrática, y, con datos reales cuyo ajuste es siempre menos que perfecto, en una situación así probablemente nos hubiéramos dado por satisfechos con una función lineal que para este rango de X debe operar muy razonablemente en el pronóstico

```
 PROGRAM TERMINATED
```

2. Una función cuadrática incompleta.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
/REGRESSION
 DEPENDENT = Y2.
 INDEPENDENT = X2.
 METHOD=NONE.
/PRINT
 MATRICES=RESI.
/PLOT
 XVAR= X,  X2, Y2,   X.
 YVAR= Y2, Y2, X(5), X(6).
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

Aquí creamos una dependiente Y2 artificial "a medida" para que ajuste a una cuadrática con solo el tercer término con coeficiente 0'07

Vanos a resolver como una lineal simple sobre X2 que es X al cuadrado. La intercept (primer término) será 0, dado como hemos creado Y2.

Solicitamos que muestre los residuales.

Pedimos una variedad de gráficos., incluidos residuales y pronosticados.

OUTPUT SELECCIONADO

```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

NUMBER OF CASES READ. . . . . .      20


SUMMARY STATISTICS FOR EACH VARIABLE
------------------------------------
                              STANDARD     COEFFICIENT    SMALLEST       LARGEST
  VARIABLE      MEAN          DEVIATION    OF VARIATION    VALUE          VALUE
------------ ------------ ------------ ------------- ------------ -----------

  2 X2           26.61151     33.57645       1.261727      0.00000    100.00000
  4 Y2            1.86281      2.35035       1.261727      0.00000      7.00000
```

El programa determina correctamente que el coeficiente que multiplica a X2 es 0,07 pero comete un error de imprecisión asignando al coeficiente intercept un valor realmente muy pequeño próximo a 0.
La ecuación obtenida es pues:

$$Y2' = 0{,}07\,X^2$$

```
                                                                        CONTRI-
  VARIABLE       REGRESSION       STANDARD  STAND.      T-   2TAIL      TOL-  BUTION
NO.   NAME       COEFFICIENT         ERROR   COEF.    STAT.    SIG.   ERANCE TO R-SQ

   INTERCEPT 0.67345179E-8 0.22530810E-7    0.000     0.30   0.768
  2 X2           0.0700000 0.53415423E-9    1.000  1.3E+8   0.000 1.000000 0.95333
```

En los datos solo hay una variable directa u observada, X; en las transformaciones introducimos sucesivamente las variables transformadas X2, Y1 e Y2. Independientemente del nombre que se le de a las variables BMDP las llama siempre por defecto X(1), X(2), X(3), X(4), tanto a observadas como a transformadas por el orden de creación. De modo que, por ejemplo podríamos referirnos a Y2 indistintamente como Y2 (el nombre que le hemos dado) o como X(4). Al solicitar que muestre los residuales 9R ofrece la matriz siguiente, donde están los datos de la variable dependiente (aquí Y2, ó X(4), los valores pronosticados para Y2 por la ecuación calculada, (que pueden referenciarse en este programa como X(5), dado que hay hasta X(4) entre variables directas y transformadas), el error típico de los valores pronosticados, los residuales (que pueden referenciarse como X(6)) y otros indicadores relacionados con los residuales (el listado continua en la página siguiente, aunque

```
                                  STANDARD
 CASE    CASE    OBSERVED  PREDICTED ERROR OF                        WEIGHTED
LABEL     NO.    Y2          VALUE  PRED.VAL.  RESIDUAL   WEIGHT    RESIDUAL
-----  -------  ---------- --------- --------- ---------  ------   ---------
            1      7.0000    7.0000    0.0000    0.0000   1.000      0.0000
            2      0.6300    0.6300    0.0000    0.0000   1.000      0.0000
            3      0.2800    0.2800    0.0000    0.0000   1.000      0.0000
            4      1.1200    1.1200    0.0000    0.0000   1.000      0.0000
            5      1.7500    1.7500    0.0000    0.0000   1.000      0.0000
            6      3.4300    3.4300    0.0000    0.0000   1.000      0.0000
            7      2.6047    2.6047    0.0000    0.0000   1.000      0.0000
            8      0.0000    0.0000    0.0000    0.0000   1.000      0.0000
            9      0.0028    0.0028    0.0000    0.0000   1.000      0.0000
           10      0.1575    0.1575    0.0000    0.0000   1.000      0.0000
           11      7.0000    7.0000    0.0000    0.0000   1.000      0.0000
           12      0.0175    0.0175    0.0000    0.0000   1.000      0.0000
           13      0.0343    0.0343    0.0000    0.0000   1.000      0.0000
           14      0.0000    0.0000    0.0000    0.0000   1.000      0.0000
           15      0.8575    0.8575    0.0000    0.0000   1.000      0.0000
           16      2.1952    2.1952    0.0000    0.0000   1.000      0.0000
           17      4.0432    4.0432    0.0000    0.0000   1.000      0.0000
           18      5.5447    5.5447    0.0000    0.0000   1.000      0.0000
           19      0.5887    0.5887    0.0000    0.0000   1.000      0.0000
           20      0.0000    0.0000    0.0000    0.0000   1.000      0.0000
```

CASE LABEL	CASE NO.	STAND-ARDIZED RESIDUAL	DELETED (PRESS) RESIDUAL	ADJUSTED (PRESS) PRED.VAL.	MAHALA-NOBIS DISTANCE	COOK DISTANCE
	1	0.34	0.0000	7.0000	4.78	0.03
	2	-0.12	0.0000	0.6300	0.28	0.00
	3	-0.06	0.0000	0.2800	0.45	0.00
	4	0.04	0.0000	1.1200	0.10	0.00
	5	0.01	0.0000	1.7500	0.00	0.00
	6	0.99	0.0000	3.4300	0.44	0.04
	7	-1.08	0.0000	2.6047	0.10	0.03
	8	-0.09	0.0000	0.0000	0.63	0.00
	9	-0.09	0.0000	0.0028	0.63	0.00
	10	-0.10	0.0000	0.1575	0.53	0.00
	11	0.34	0.0000	7.0000	4.78	0.03
	12	-0.09	0.0000	0.0175	0.62	0.00
	13	-0.08	0.0000	0.0343	0.61	0.00
	14	-0.09	0.0000	0.0000	0.63	0.00
	15	0.18	0.0000	0.8575	0.18	0.00
	16	0.78	0.0000	2.1952	0.02	0.02
	17	1.93	0.0000	4.0432	0.86	0.20
	18	-2.73	0.0000	5.5447	2.45	0.81
	19	-0.06	0.0000	0.5887	0.29	0.00
	20	-0.09	0.0000	0.0000	0.63	0.00

SUMMARY STATISTICS FOR RESIDUALS

(CASES WITH POSITIVE WEIGHT)

AVERAGE RESIDUAL	0.0000
RESIDUAL MEAN SQUARE	0.00000000
AVERAGE DELETED RESIDUAL	0.0000
AVE. SQUARED DELETED RESIDUAL (PREDICTION MEAN SQUARE)	0.00000000
SERIAL CORRELATION	-0.3119
DURBIN-WATSON STATISTIC	2.6142

Diagnósticos acerca de los residuales.

4.78 IS THE MAXIMUM VALUE OF MAHALANOBIS DISTANCE AMONG CASES WITH POSITIVE CASE WEIGHT. THIS OCCURRED FOR CASE NUMBER 1
CASE LABEL =

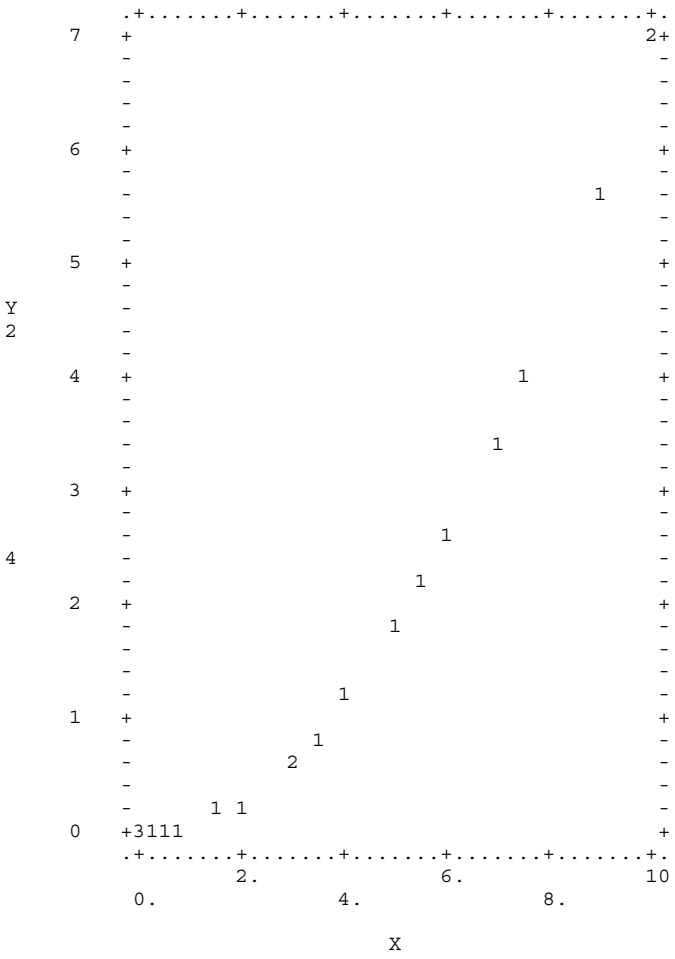
-2.73 IS THE LARGEST STANDARDIZED RESIDUAL (IN ABSOLUTE VALUE) AMONG CASES WITH POSITIVE CASE WEIGHT.
THIS OCCURRED FOR CASE NUMBER 18, CASE LABEL =

0.81 IS THE MAXIMUM VALUE OF COOK'S DISTANCE AMONG CASES WITH POSITIVE WEIGHT. THIS OCCURRED FOR CASE NUMBER 18
CASE LABEL =
IF THIS CASE WERE OMITTED, THE REGRESSION COEFFICIENTS WOULD MOVE FROM THE VALUES REPORTED ABOVE TO THE EDGE OF A 62.11 PERCENT CONFIDENCE ELLIPSOID.

COMPARISON OF ESTIMATES OF REGRESSION COEFFICIENTS
(RELATIVE DIFFERENCE IS DIFFERENCE DIVIDED BY ORDINARY COEF.
STANDARD ERROR IS THAT OF ORDINARY COEFFICIENT.)

	ORDINARY LEAST SQUARES	OMITTING CASE WITH LARGEST COOK DISTANCE	RELATIVE DIFFERENCE	DIFFERENCE DIVIDED BY STANDARD ERROR
INTERCEPT	0.000000	0.000000	0.5370	0.1605
2 X2	0.070000	0.070000	0.0000	-1.0833

Gráficos:



Variable dependiente (Y2) en ordenadas por independiente (X) en abcisas.

En este caso se aprecia muy bien la curvatura que reclama ensayar una función cuadrática.

Compárese con el gráfico siguiente donde en abcisas hemos colocado la transformada X2 que es X elevado al cuadrado.

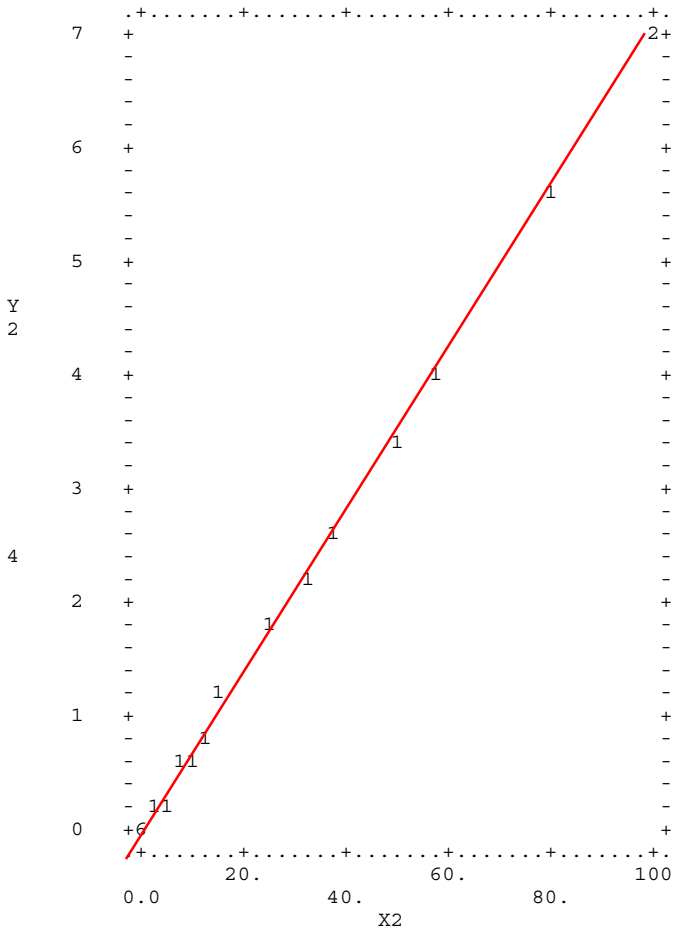
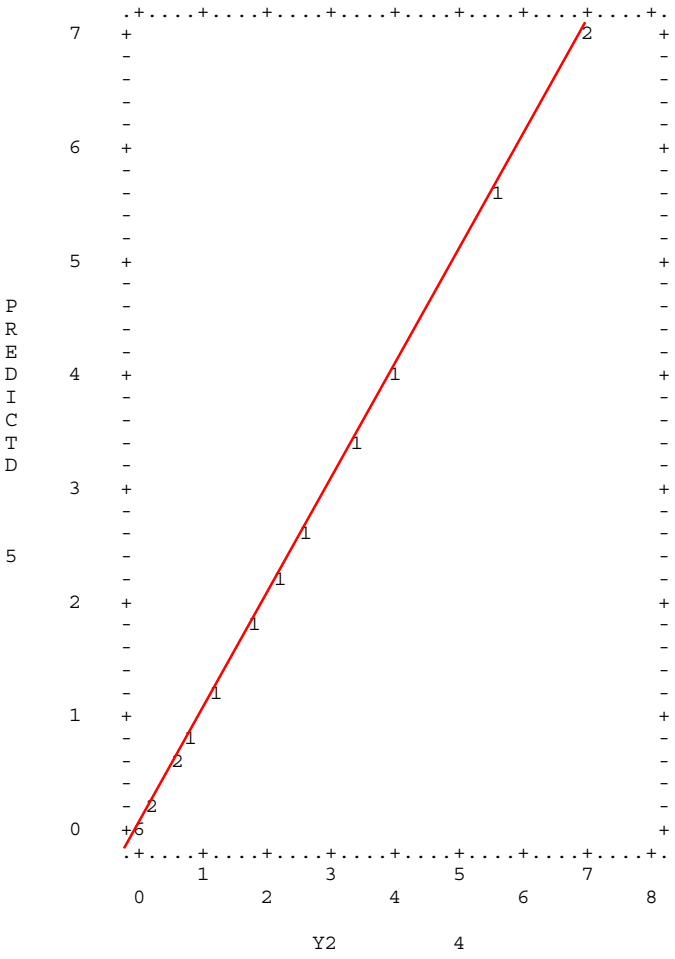


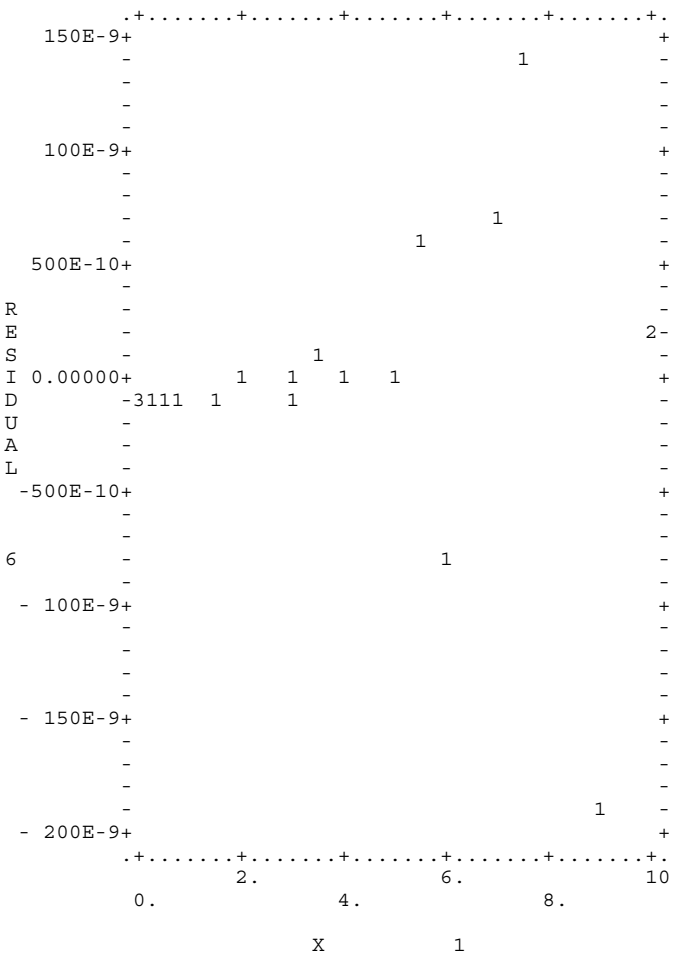
Diagrama de dispersión X2 versus Y2.

En ordenadas la dependiente Y2; en abcisas la transformada X2 que hemos creado para linearizar la relación entre Y2 y X y poder tratarla mediante regresión lineal.
Obsérvese que, efectivamente, Y2 frente a X2 presenta una clara recta.
Esta reducción del brazo de parábola del plot anterior a una recta es señal clara de que hemos transformado de un modo adecuado para emprender un tratamiento lineal de la relación. La transformada X2 (que contiene X al cuadrado) puede ahora usarse en una ecuación lineal para pronosticar Y2, porque X2 mantiene una relación lineal



Aquí tenemos un gráfico de los valores "reales" de Y2 (en abcisas) frentes a los valores pronosticados por la función para Y2 (en ordenadas).
Como el pronóstico es perfecto (construimos artificialmente Y2 para cumplir este requisito) estos pares describen perfectamente una recta.

Hemos conseguido este gráfico llamando a los valores pronosticados en las instrucciones por su nombre de variable por defecto que es X(5).
Es X(5) en este caso porque entre variables directas y transformadas hay definidas 4, por tanto en 9R la variable "valores pronosticados para la variable dependiente" es la número 5. (Siempre es la primera siguiente a las



Residuales, que son la variable X(6), frente a la variable independiente X.

Aquí todos los residuales deberían ser 0. No son exactamente 0, sino valores muy cercanos a 0 (como se ve) por inexactitud acumulada a partir del sexto decimal. La variabilidad que muestra el gráfico es fruto de imprecisión y no de verdaderos residuales. Quizás esta imprecisión es debida a que en los cálculos no críticos el programa trabaja en simple precisión y al celo del programa por mostrar residuales muy pequeños, hasta 9 y 10 lugares a la derecha de la coma decimal. En cualquier caso el residual más grande que vemos en el gráfico es tan pequeño que, aunque no supiéramos de entrada que el ajuste es en realidad perfecto, seguiríamos considerando que el ajuste es prácticamente perfecto.

En una situación de regresión sobre una dependiente real este diagrama tiene interés para ver si el error varía

PROGRAM TERMINATED

3. Una función polinómica de tercer grado.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 X3 = X**3.
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
 Y3 = -107 + 1.3*X - X2 + 0.9*X3.
```

Transformadas necesarias para linearizar los términos de la función cubica y resolverla como una función lineal.

Generación de una dependiente "ad hoc" que estamos seguros que ajusta a una función polinómica de tercer grado.

```
/REGRESSION
 DEPENDENT = Y3.
 INDEPENDENT = X, X2, X3.
 METHOD=NONE.
```

Planteamiento de la resolución de la función. Y3 es la dependiente a pronosticar. X, X2 y X3 entran como tres variables independientes.
Con METHOD=NONE se desactiva la búsqueda de mejores predictores y el programa introduce todas las variables independientes en la ecuación.

```
/PRINT
 CASE=20.
 MATRICES=RESI.
/PLOT
 XVAR= X, X2, X3.
 YVAR=Y3, Y3, Y3.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

OUTPUT SELECCIONADO

BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

Datos, estadísticos, primer dígito de correlaciones y estadísticos para el subconjunto.

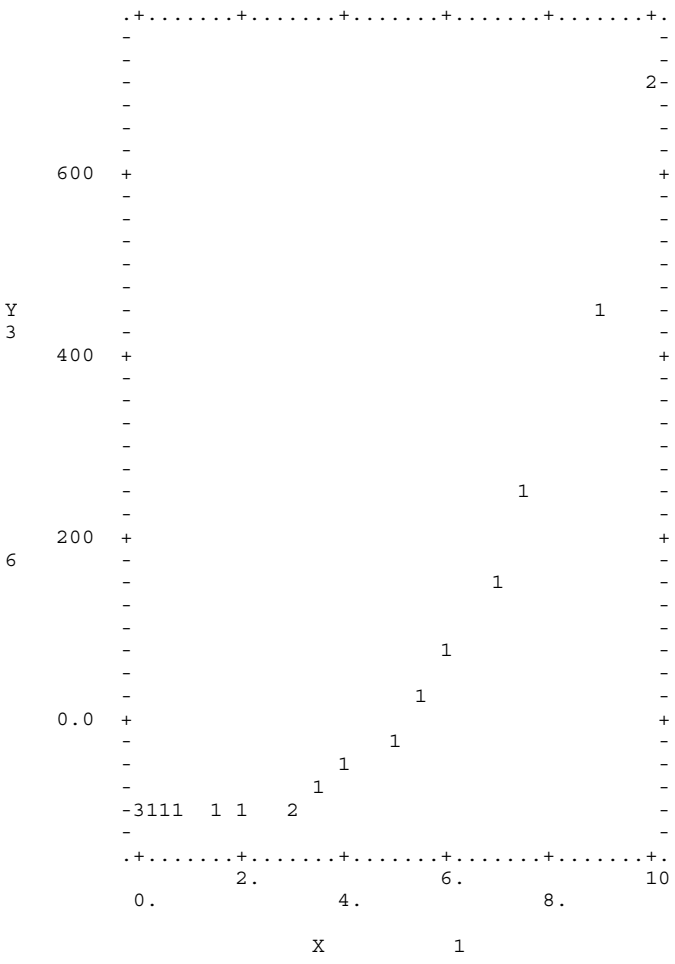
VARIABLE NO.	NAME	REGRESSION COEFFICIENT	STANDARD ERROR	STAND. COEF.	T-STAT.	2TAIL SIG.	TOLERANCE	CONTRIBUTION TO R-SQ
	INTERCEPT	-107.000	0.29916456E-5	-0.403	-3.6E+7	0.000		
1	X	1.30000	0.34536850E-5	0.017	3.8E+5	0.000	0.015609	0.00000
2	X2	-1.00000	0.90615760E-6	-0.126	-1.1E+6	0.000	0.002371	0.00003
3	X3	0.900000	0.61231333E-7	1.109	1.5E+7	0.000	0.005461	0.00600

Solución de la Ecuación:

$$Y3' = -107 + 1{,}3X - X^2 + 0{,}9X^3$$

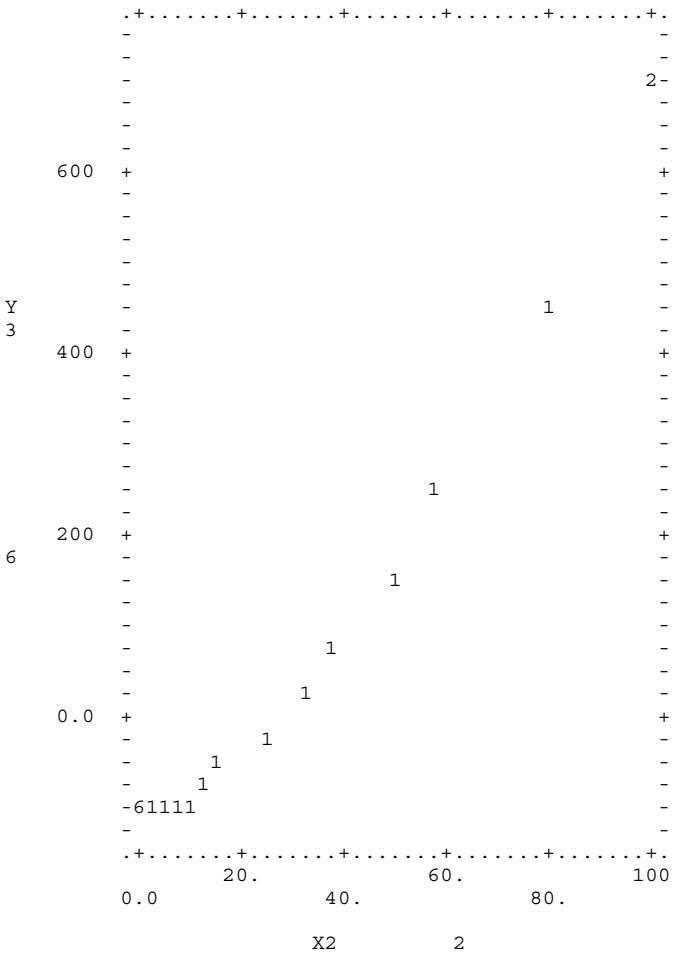
CASE LABEL	CASE NO.	OBSERVED Y3	PREDICTED VALUE
	1	706.0000	706.0000
	2	-87.8000	-87.8000
	3	-101.2000	-101.2000
	4	-60.2000	-60.2000
	5	-13.0000	-13.0000
	6	161.8000	161.8000
	7	68.0029	68.0029
	8	-106.9871	-106.9871
	9	-106.7728	-106.7728
	10	-104.2625	-104.2625
	11	706.0000	706.0000
	12	-106.4875	-106.4875
	13	-106.2713	-106.2713
	14	-106.9948	-106.9948
	15	-76.1125	-76.1125
	16	26.9744	26.9744
	17	240.1984	240.1984
	18	459.8321	459.8320
	19	-89.6899	-89.6899
	20	-106.9997	-106.9997

Valores observados en la variable dependiente Y3 y valores estimados o pronosticados para la variable dependiente, es decir, Y3'



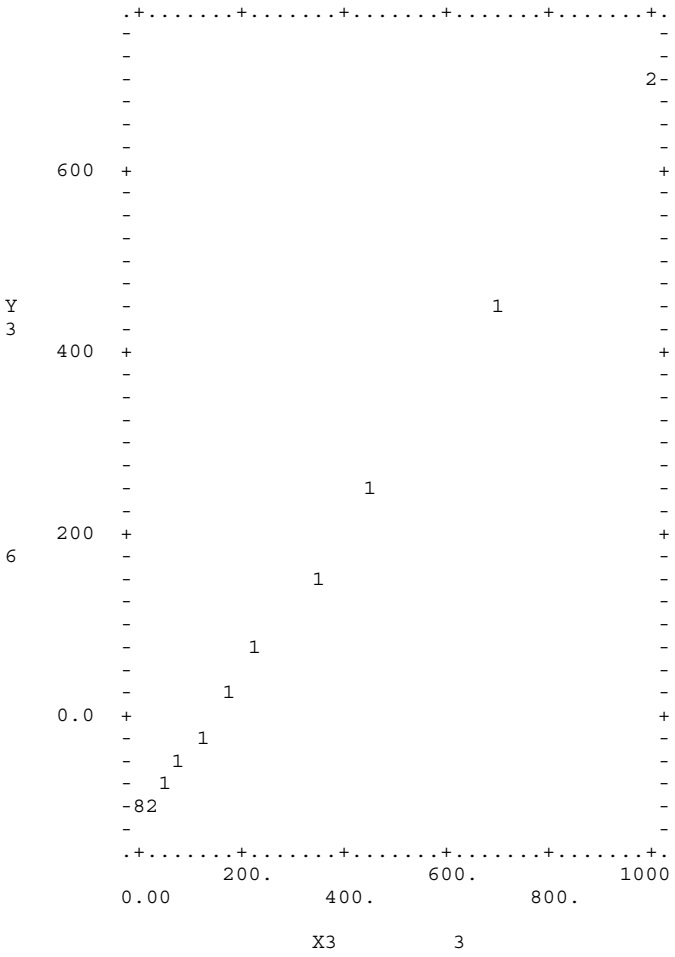
La variable dependiente Y3 frente a X.

La curva sugiere inmediatamente una función no lineal.



La dependiente Y3 frente a X elevado al cuadrado.

Puede apreciarse que la relación no se ha linearizado perfectamente, lo que sugiere ensayar una función con un grado más alto.



La variable dependiente Y3 frente a la variable X elevada al cubo.

Ahora la relación es prácticamente lineal a simple vista, esto sugiere ensayar una función polinómica de tercer grado

Obsérvese que si esta función polinómica de tercer grado fuera incompleta, (es decir los coeficientes para X ó para X al cuadrado, o para ambos, fueran 0), un cálculo suficientemente preciso lo pondría de manifiesto al resolver como una polinómica de grado 3

PROGRAM TERMINATED

4. Una función polinómica incompleta de cuarto grado tratada como completa.

En el siguiente ejemplo se va resolver una función polinómica de cuarto grado incompleta. Primero se resolverá tratándola como si fuera completa para apreciar como los resultados nos conducirían a recalcularla como incompleta.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 X3 = X**3.
 X4 = X**4.
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
 Y3 = -107 + 1.3*X - X2 + 0.9*X3.
 Y4 = -0.07*X4.
/REGRESSION
 DEPENDENT = Y4.
 INDEPENDENT = X, X2, X3, X4.
 METHOD=NONE.
/PRINT
 CASE=20.
 MATRICES=RESI.
/PLOT
 XVAR= X.
 YVAR=Y4.
/END
```

Creación de los términos independientes para linearizar la función.

Creación de la variable dependiente.

Planteamiento de la ecuación.

OUTPUT SELECCIONADO

(Invertimos el orden de presentación del output seleccionado).

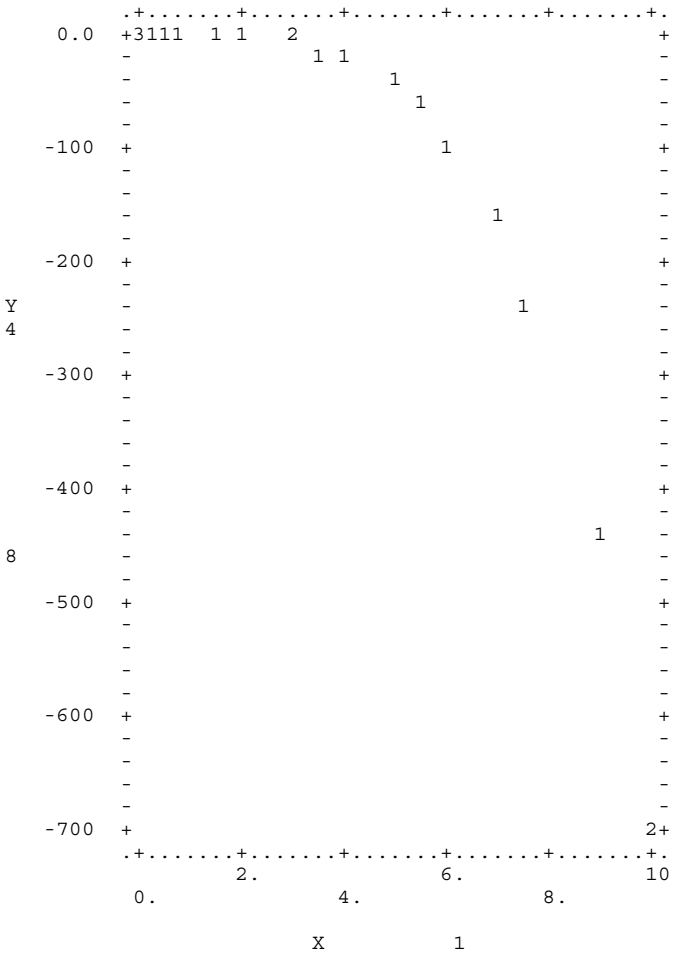


Diagrama de dispersión de la variable dependiente Y4 (en ordenadas) frente a la variable independiente X (en abcisas).

Los datos describen claramente una curva que decrece progresivamente desde Y4 igual a 0 aproximadamente. Esto sugiere dos cosas. Primero, que la relación es no-lineal. Segundo que si se trata de una función polinómica probablemente al menos el coeficiente de más grado (que a igualdad de pesos es el que marca la parte principal de cada valor pronosticado) tiene signo

```
VARIABLE    1 X        IS NOT INCLUDED AS A PREDICTOR DUE TO LOW TOLERANCE.
                                                                      CONTRI-
  VARIABLE      REGRESSION      STANDARD  STAND.     T-  2TAIL     TOL-  BUTION
NO.   NAME     COEFFICIENT         ERROR   COEF.  STAT.   SIG.   ERANCE TO R-SQ

   INTERCEPT-0.76267251E-7 0.17245936E-5  -0.000  -0.04  0.965
  2 X2       0.19947051E-6 0.61251905E-6   0.000   0.33  0.749 0.002479 0.00000
  3 X3      -0.73807747E-7 0.15127436E-6  -0.000  -0.49  0.632 0.000427 0.00000
  4 X4          -0.0700000 0.93298364E-8  -1.000-7.5E+6  0.000 0.001166 0.00117

NOTE THAT    1 VARIABLES HAVE BEEN OMITTED BECAUSE OF LOW TOLERANCE.
```

Los resultados muestran que: Primero, la variable X está tan relacionada con estos predictores que si aspiramos a incluirla en la ecuación hay que bajar el límite por defecto de la tolerancia. Por ejemplo, mediante la instrucción TOL=0.0l el nivel de tolerance puede fijarse a 0,01. Por defecto 9R trabaja con TOL=0.0001 que es una tolerancia que permite ya una fuerte colinealidad. Sin embargo, cuando METHOD=NONE, como en este caso, se calcula la regresión para un conjunto de variables no redundantes. Segundo, los coeficiente de los términos X2, X3 e intercept son tan pequeños (nótese que están en notación científica) que este resultado sugiere recalcular la ecuación prescindiendo de ellos. Es decir, el resultado sugiere estimar la ecuación incompleta con solo X4.

5. Una función polinómica incompleta de cuarto grado.

En función de los resultados anteriores en el siguiente ejemplo se va a resolver la función polinómica de cuarto grado como incompleta.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 X3 = X**3.
 X4 = X**4.
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
 Y3 = -107 + 1.3*X - X2 + 0.9*X3.
 Y4 = -0.07*X4.
/REGRESSION
 DEPENDENT = Y4.
 INDEPENDENT = X4.
 METHOD=NONE.
/PRINT
 CASE=20.
 MATRICES=RESI.
/PLOT
 XVAR= X.
 YVAR=Y4.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

La ecuación se plantea ahora solo con el término de cuarto grado.

OUTPUT SELECCIONADO

```
                                                                      CONTRI-
  VARIABLE      REGRESSION      STANDARD  STAND.     T-  2TAIL     TOL-  BUTION
NO.   NAME     COEFFICIENT         ERROR   COEF.  STAT.   SIG.   ERANCE TO R-SQ

   INTERCEPT-0.48200139E-6 0.11161531E-5  -0.000  -0.43  0.671
  4 X4          -0.0700000 0.30978223E-9  -1.000-2.3E+8  0.000 1.000000 0.94480
```

La intercept es prácticamente 0 y la ecuación puede escribirse así:

$$Y4' = -0.07\,X^4$$

El ajuste de la ecuación (no mostrado) es perfecto.

6. Una función potencial.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 X3 = X**3.
 X4 = X**4.
 XLOG = LOG(X).
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
 Y3 = -107 + 1.3*X - X2 + 0.9*X3.
 Y4 = -0.07*X4.
 Y5 = 7*(X**2).
 Y5LOG = LOG(Y5).
/REGRESSION
 DEPENDENT = Y5LOG.
 INDEPENDENT = XLOG.
 METHOD=NONE.
/PRINT
 CASE=20.
/PLOT
 XVAR=XLOG.
 YVAR=Y5LOG.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

Con propósitos didácticos, se crea una dependiente Y5 que es función potencial de X.

Para resolver necesitamos una variable que represente el logaritmo de X, que llamaremos XLOG, y otra que represente el logaritmo de Y5, que llamaremos Y5LOG.

Se utilizan logaritmos decimales, pero en realidad la base de logaritmación es indiferente para estos cálculos con tal que se tenga en cuenta posteriormente cuando se requieran antilogaritmos.

La ecuación a resolver coloca a Y5LOG como dependiente y a XLOG como independiente.

OUTPUT SELECCIONADO

```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

STATISTICS FOR 'BEST' SUBSET
-----------------------------
SQUARED MULTIPLE CORRELATION    1.00000
MULTIPLE CORRELATION            1.00000
ADJUSTED SQUARED MULT. CORR.    1.00000
RESIDUAL MEAN SQUARE           0.000000
STANDARD ERROR OF EST.         0.000000
F-STATISTIC                  **********
NUMERATOR DEGREES OF FREEDOM          1
DENOMINATOR DEGREES OF FREEDOM       18
SIGNIFICANCE (TAIL PROB.)        0.0000


----------------------------------------

                                                                    CONTRI-
  VARIABLE       REGRESSION       STANDARD  STAND.      T-   2TAIL     TOL-   BUTION
NO.   NAME      COEFFICIENT          ERROR   COEF.    STAT.    SIG.   ERANCE TO R-SQ

   INTERCEPT      0.845098 0.21369884E-7    0.329 4.0E+7   0.000
   5 XLOG         2.00000  0.17078512E-7    1.000 1.2E+8   0.000 1.000000 1.01503
```

El ajuste de la ecuación a los datos es perfecto.

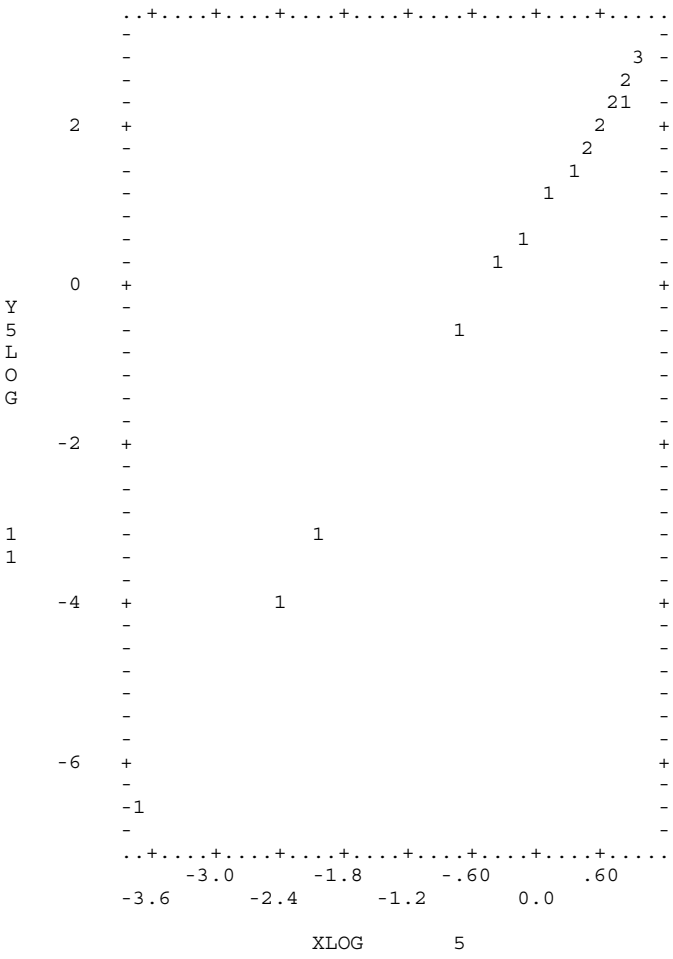
El resultado obtenido puede escribirse

$$\log(Y5)' = 0{,}845098 + 2 \cdot \log(X).$$

En ese resultado se dispone del valor de b=2 en la ecuación potencial, pero necesitaremos obtener el antilogaritmo decimal de a = 0,845098 para poder escribir la ecuación en forma de ecuación potencial. El antilog(0,845098)=7, por tanto la ecuación potencial que relaciona Y5 y X es:

Antilog(0,845098)=7

$$Y5' = 7 \cdot X^2$$



El diagrama de dispersión adjunto muestra que la relación entre las transformadas de la dependiente y de la independiente presentan una relación lineal.

7. Una función exponencial.

```
INSTRUCCIONES. PROGRAMA 9R.
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 X3 = X**3.
 X4 = X**4.
 XLOG = LOG(X).
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
 Y3 = -107 + 1.3*X - X2 + 0.9*X3.
 Y4 = -0.07*X4.
 Y5 = 7*(X**2).
 Y5LOG = LOG(Y5).
 Y6 = 3*(2**X).
 Y6LOG = LOG(Y6).

/REGRESSION
 DEPENDENT = Y6LOG.
 INDEPENDENT = X.
 METHOD=NONE.
/PRINT
 CASE=20.
/PLOT
 XVAR=X, X.
 YVAR=Y6, Y6LOG.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

La exponencial solo requiere transformar la dependiente en su logaritmo. Creamos una dependiente Y6 como función exponencial de X y a continuación creamos una transformada de la misma Y6LOG que contiene su logaritmo decimal.
En la ecuación X entra sin transformar como independiente.

OUTPUT SELECCIONADO

```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

NUMBER OF CASES READ. . . . . .      20

CORRELATIONS
              X        Y6LOG
                1       13

X           1    1.000
Y6LOG      13    1.000    1.000

STATISTICS FOR 'BEST' SUBSET
----------------------------
SQUARED MULTIPLE CORRELATION    1.00000
MULTIPLE CORRELATION            1.00000
ADJUSTED SQUARED MULT. CORR.    1.00000
RESIDUAL MEAN SQUARE           0.000000
STANDARD ERROR OF EST.         0.000000
F-STATISTIC                  **********
NUMERATOR DEGREES OF FREEDOM          1
DENOMINATOR DEGREES OF FREEDOM       18
SIGNIFICANCE (TAIL PROB.)        0.0000

                                                                    CONTRI-
  VARIABLE       REGRESSION       STANDARD  STAND.      T-   2TAIL      TOL-   BUTION
NO.   NAME      COEFFICIENT          ERROR   COEF.   STAT.    SIG.    ERANCE TO R-SQ

   INTERCEPT       0.477121 0.16471045E-7    0.462 2.9E+7   0.000
  1 X              0.301030 0.31929083E-8    1.000 9.4E+7   0.000 1.000000 0.98686
```

El ajuste es perfecto, fruto de la creación de los valores de la dependiente como transformación de la independiente.

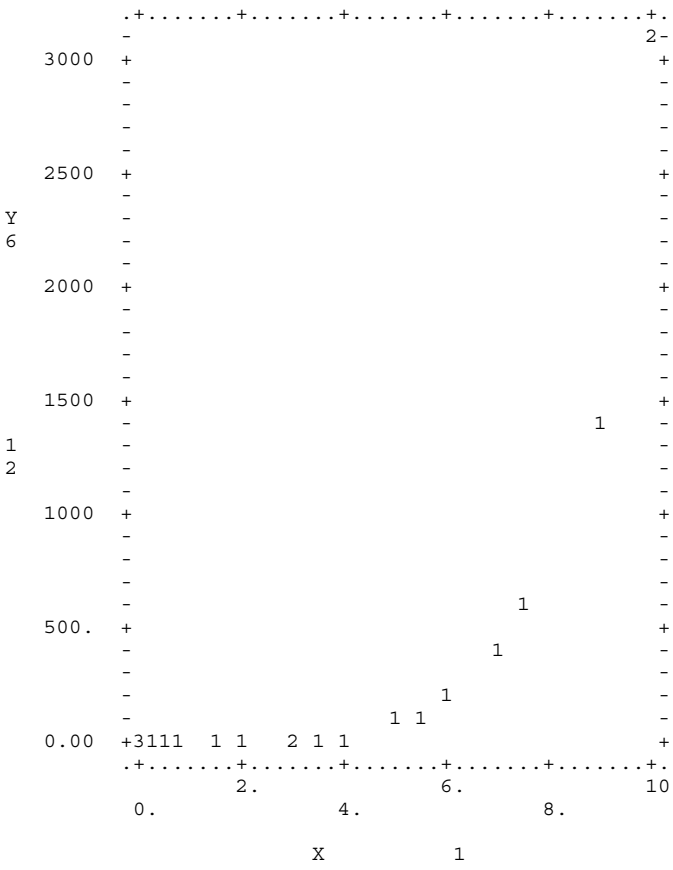
$$\log(Y6)' = 0{,}477121 + 0{,}301030\,X$$

Antilog(0,477121)=3

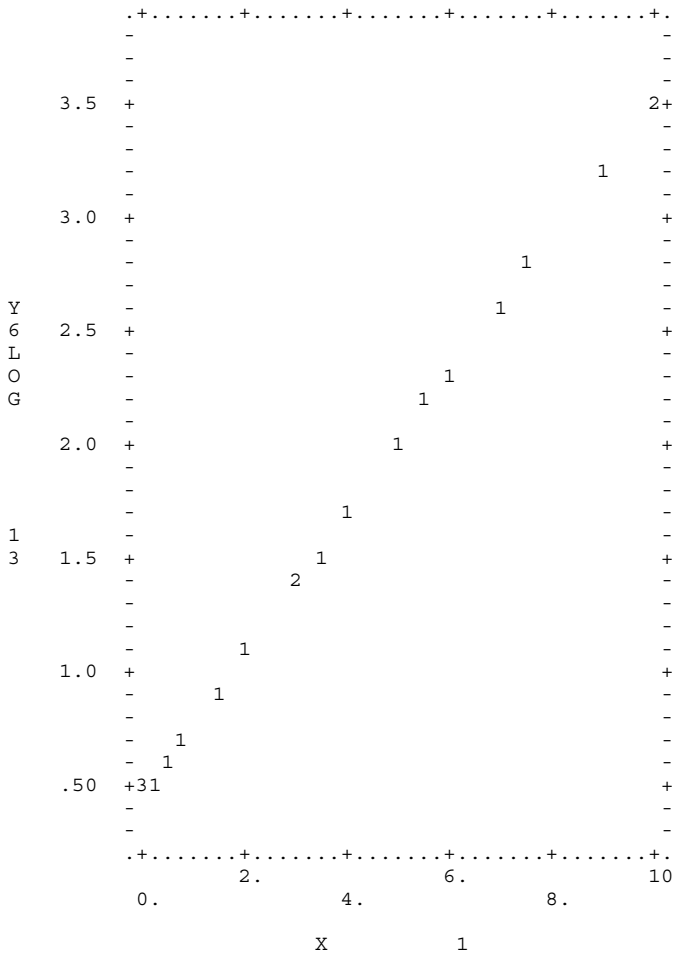
Antilog(0,301030)=2

$$Y6' = 3 \cdot 2^{X}$$

Después de obtener los antilogaritmos decimales de los dos coeficientes podemos escribir la función de regresión en formato exponencial.



Y6 frente a X. La presencia de una relación no-lineal bien formada es evidente, aunque puede ser difícil decidir exactamente la forma de la función con la mera inspección visual del gráfico, particularmente con datos reales donde si que hay error presente



La transformación de la variable dependiente a su logaritmo ha bastado para obtener una relación manifiestamente lineal: esto puede interpretarse como un indicador de que estamos ante una función de tipo exponencial.

8. Una función logarítmica.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 X2 = X**2.
 X3 = X**3.
 X4 = X**4.
 XLOG = LOG(X).
 Y1 = 13.5 + 14*X - 0.5*X2.
 Y2 = 0.07*X2.
 Y3 = -107 + 1.3*X - X2 + 0.9*X3.
 Y4 = -0.07*X4.
 Y5 = 7*(X**2).
 Y5LOG = LOG(Y5).
 Y6 = 3*(2**X).
 Y6LOG = LOG(Y6).
 Y7 = 2 + 3*XLOG.
/REGRESSION
 DEPENDENT = Y7.
 INDEPENDENT = XLOG.
 METHOD=NONE.
/PRINT
 CASE=20.
/PLOT
 XVAR=X, XLOG.
 YVAR=Y7, Y7.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

Para la función logarítmica solo necesitamos el logaritmo de X que representamos por XLOG. Todas las demás transformadas utilizadas en otros cálculos no son necesarias aquí. Las hemos ido acumulando para mostrar como se puede ir trabajando sobre el mismo archivo de instrucciones añadiendo las transformadas que resulte necesario sin que ello afecte a los nuevos resultados.

Creamos una dependiente Y7 que es función logarítmica de X, para obtener después un ajuste perfecto. Los coeficientes serán pues 2 (intercept) y 3 para b.

En la ecuación solo hay una independiente: la transformada que recoge el logaritmo de X.

OUTPUT SELECCIONADO

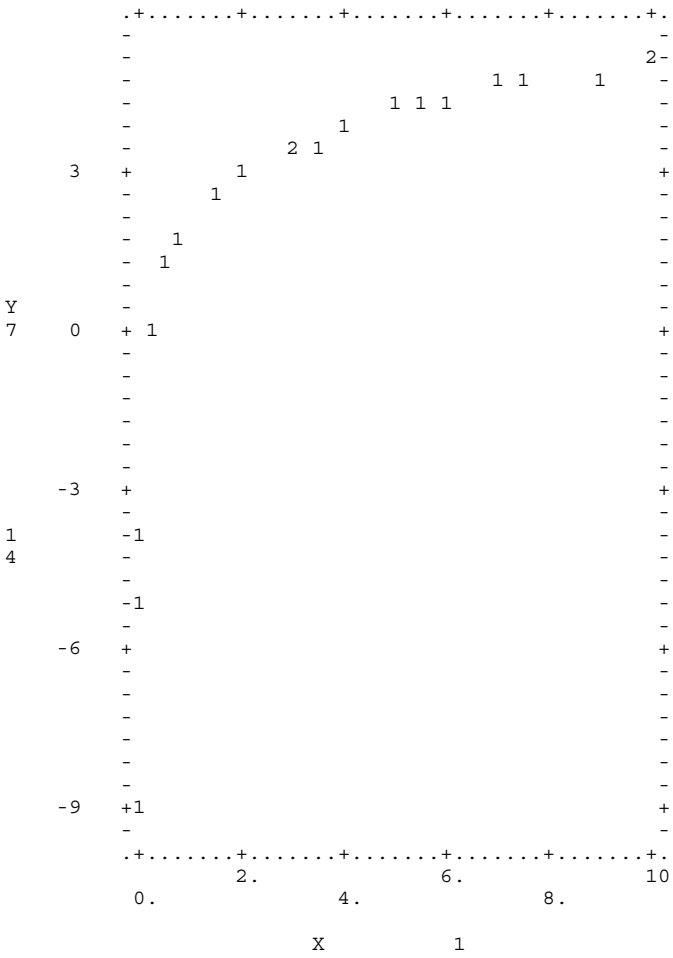
```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

SQUARED MULTIPLE CORRELATION    1.00000
ADJUSTED SQUARED MULT. CORR.    1.00000

                                                                    CONTRI-
  VARIABLE      REGRESSION       STANDARD  STAND.      T-   2TAIL     TOL-   BUTION
NO.   NAME     COEFFICIENT          ERROR   COEF.   STAT.    SIG.   ERANCE TO R-SQ

   INTERCEPT       2.00000 0.20664121E-7   0.519 9.7E+7   0.000
 5 XLOG            3.00000 0.16514476E-7   1.000 1.8E+8   0.000 1.000000 1.01770
```

El ajuste es perfecto.

La ecuación logarítmica es la siguiente:

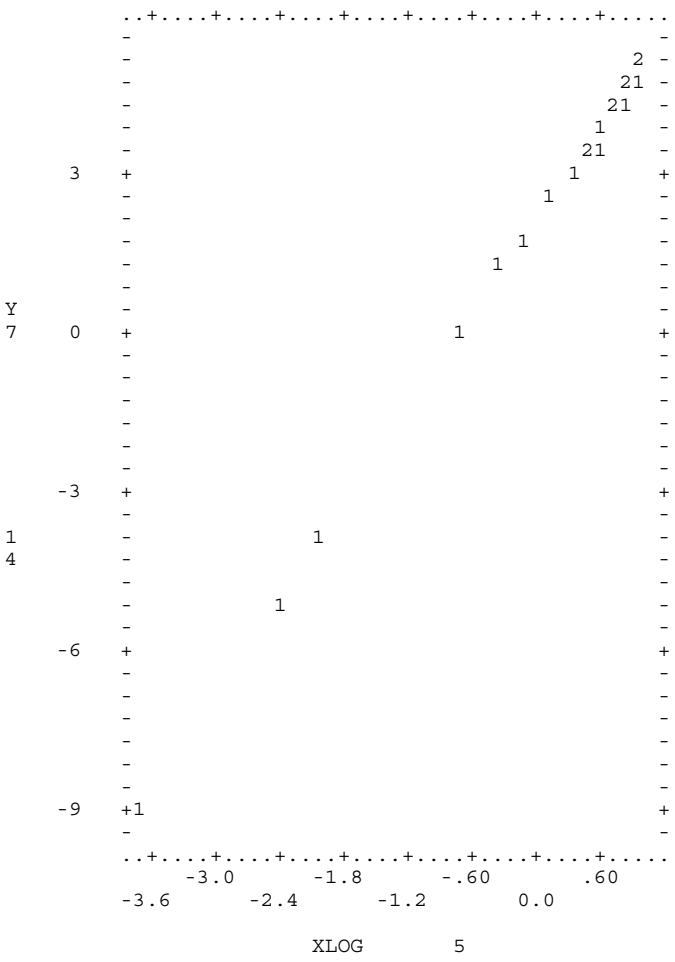
$$Y7' = 2 + 3 \cdot \log(X)$$

En este caso los coeficientes obtenidos son directamente los que entran en la ecuación.



La gráfica muestra la relación no-lineal entre Y7 y X.

Y7 es el log(X) por 3, más 2, aunque esto solo es fácilmente contrastable en el gráfico después de conocer que esos son los coeficientes y esa la forma



La relación entre log(X) e Y7 aparece lineal, lo que sugiere de inmediato ensayar una función logarítmica.

9. Una función hiperbólica.

INSTRUCCIONES. PROGRAMA 9R.

```
/INPUT TITLE IS `REGRESION NO LINEAL RESUELTA POR REGRESION LINEAL'.
 VARIABLES ARE 1. CASES ARE 20. FORMAT IS STREAM.
/VARIABLE NAME IS X.
/TRANSFORMATIONS
 XI = 1/X.
 Y8 = 120 + 17/X.
/REGRESSION
 DEPENDENT = Y8.
 INDEPENDENT = XI.
 METHOD=NONE.
/END
 10 3 2 4 5 7 6.1 0.01 0.2 1.5 10 0.5 0.7 0.004 3.5 5.6 7.6 8.9 2.9 0.0002
/END
```

Creamos una variable Y8 como función hiperbólica de X, con parámetros 120 (intercept) y 17 (b).
La variable XI contiene los recíprocos de X, necesarios para linearizar la solución.

OUTPUT SELECCIONADO

```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION


  VARIABLE      REGRESSION      STANDARD  STAND.     T-   2TAIL     TOL-  BUTION
NO.   NAME     COEFFICIENT         ERROR   COEF.   STAT.   SIG.   ERANCE TO R-SQ

   INTERCEPT       120.000  0.0000574105   0.006 2.1E+6  0.000
 15 XI             17.0000 0.51275203E-7   1.000 3.3E+8  0.000 1.000000 0.67799
```

La ecuación hiperbólica estimada es:

$$Y8' = 120 + 17 \cdot \frac{1}{X}$$

10. Diagnóstico gráfico del tipo de función basado en transformadas.

En los análisis anteriores puede observarse que los gráficos de dispersión entre variable independiente y dependiente o sus transformadas puede ayudar a determinar el tipo de función. En la práctica es posible que se desconozca a priori cual es el tipo de función que resulta más adecuada para unos datos, tarea para la que el diagnóstico gráfico puede ser de gran ayuda. Para evaluar algunas de las funciones principales mediante análisis gráfico basta con obtener ciertas variables transformadas de la variable independiente X y de la dependiente Y y ensayar los diagramas de dispersión siguientes. Si uno de ellos muestra una relación aproximadamente lineal ello indica que ese tipo de función puede ser el indicado, según la siguiente tabla.

DIAGRAMA DE DISPERSION: INDEPENDIENTE		DEPENDIENTE		TIPO DE FUNCION A UTILIZAR:
Si X	versus	Y	es aproximadamente lineal	LINEAL
" X^2	"	Y	"	CUADRATICA
" X^3	"	Y	"	POLINOMICA DE TERCER GRADO.
" X^4	"	Y	"	POLINOMICA DE CUARTO GRADO.
" LOG(X)	"	LOG(Y)	"	POTENCIAL.
" X	"	LOG(Y)	"	EXPONENCIAL.
" LOG(X)	"	Y	"	LOGARITMICA.
" 1/X	"	Y	"	HIPERBOLICA.

El grado de ajuste que muestren distintas funciones para unos mismos datos permitirá evaluar con mayor exactitud cual de ellas responde mejor, sin que deba perderse de vista que, en general, en ausencia de otras razones teóricas o prácticas relevantes, entre dos modelos con un poder predictivo semejante es preferible el más sencillo. Por ejemplo, cuantos más parámetros presenta una función es más fácil que ajuste a los datos, sin embargo, en general, es preferible no añadir parámetros que solo incrementan el poder predictivo muy poco y, salvo diferencia clara en favor de modelos no-lineales, se prefieren en la práctica los modelos lineales considerados más sencillos y manejables.

Por supuesto no todas las relaciones entre variables son lineales y no todas las funciones no lineales son tan sencillas como las anteriores. Existen otras muchas formas de funciones no lineales que pueden ser relevantes en contextos de trabajo determinados, y diversas formas de funciones no lineales pueden aparecer combinadas entre sí. Por otra parte existen procedimientos estadísticos y programas estadísticos que permiten aproximar razonablemente el valor de los coeficientes de esas funciones no lineales, aunque la solución de resolverlas reduciéndolas a funciones lineales mediante transformaciones puede ser la más sencilla e intuitiva para estas funciones simples.