Tercera práctica de REGRESIÓN.

DATOS: fichero “practica regresión 3.sf3”

1. Objetivo:

El objetivo de esta práctica es aplicar el modelo de regresión con más de una variable explicativa. Es decir regresión múltiple.

Para ello, en la primera parte de la práctica se abordará la estimación de un modelo de regresión múltiple

En la segunda parte, se realizará la interpretación del modelo, analizando qué variables son significativas y el significado de los parámetros.

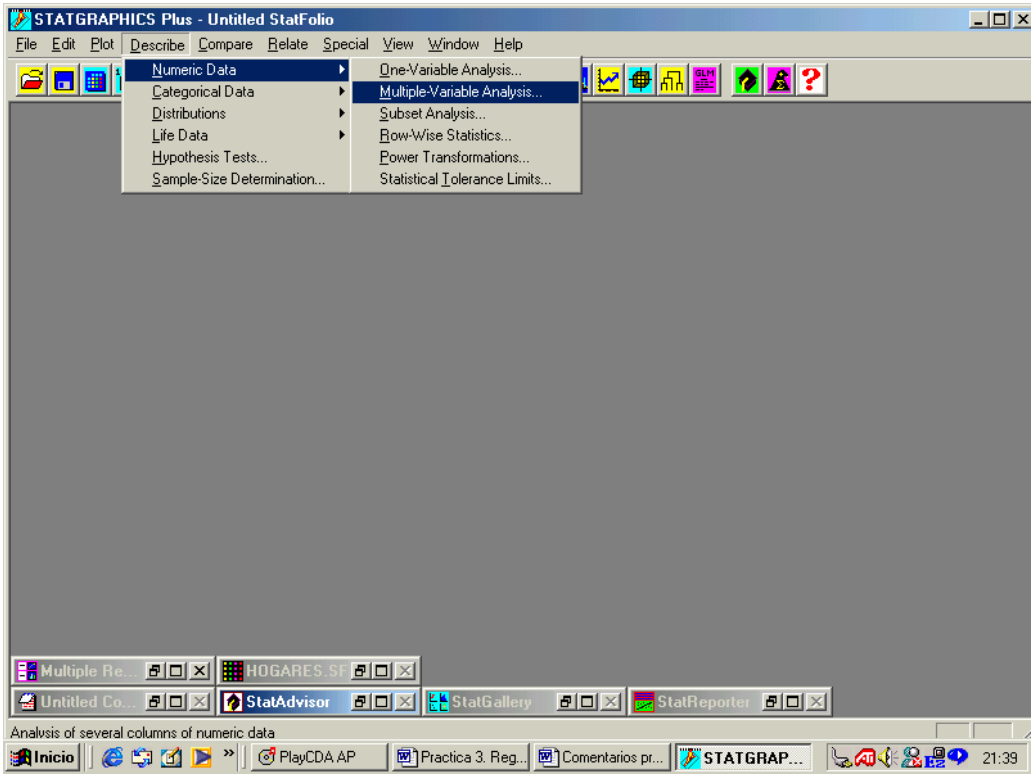
En la tercera parte, se realizará la diagnosis del modelo y se predecirán valores de la variable respuesta.

Conocimientos necesarios de otras prácticas

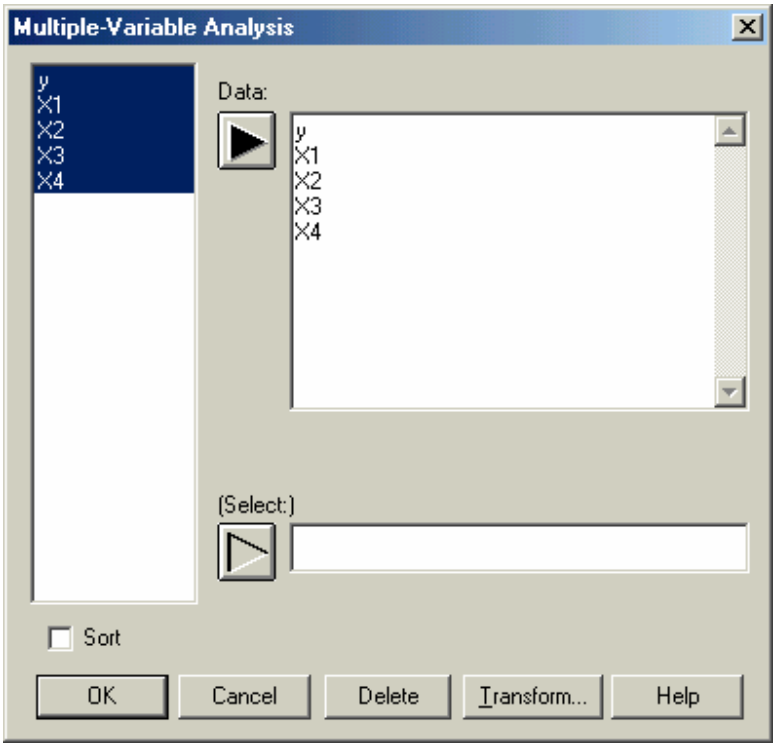
Estimación, interpretación de parámetros y diagnosis de un modelo de regresión simple. Modelos de regresión con datos transformados

2. Estimación en Regresión Múltiple.

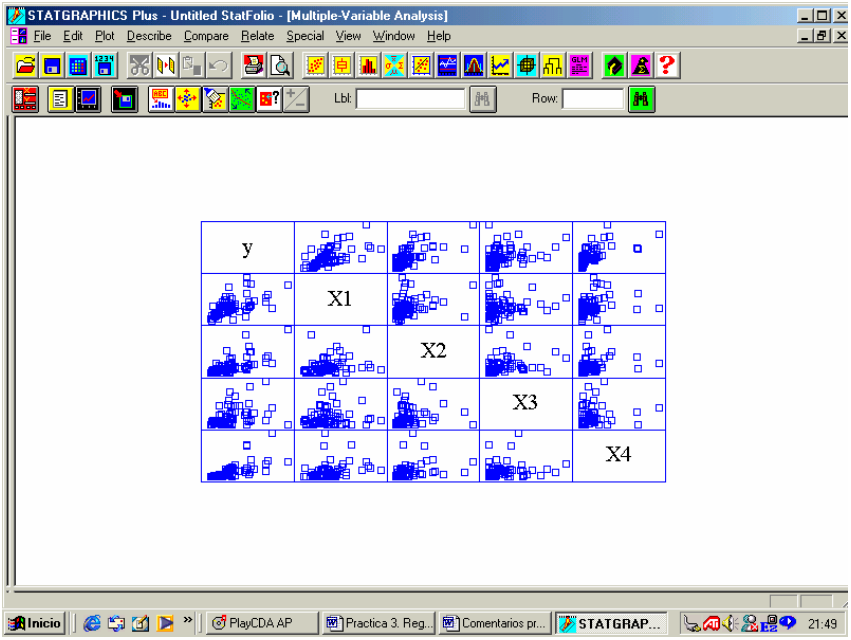
Antes de estimar el modelo de regresión múltiple comprobaremos que las relaciones entre las variables son lineales. Para ello realizaremos un grafico de dispersión múltiple entre ellas como sigue:



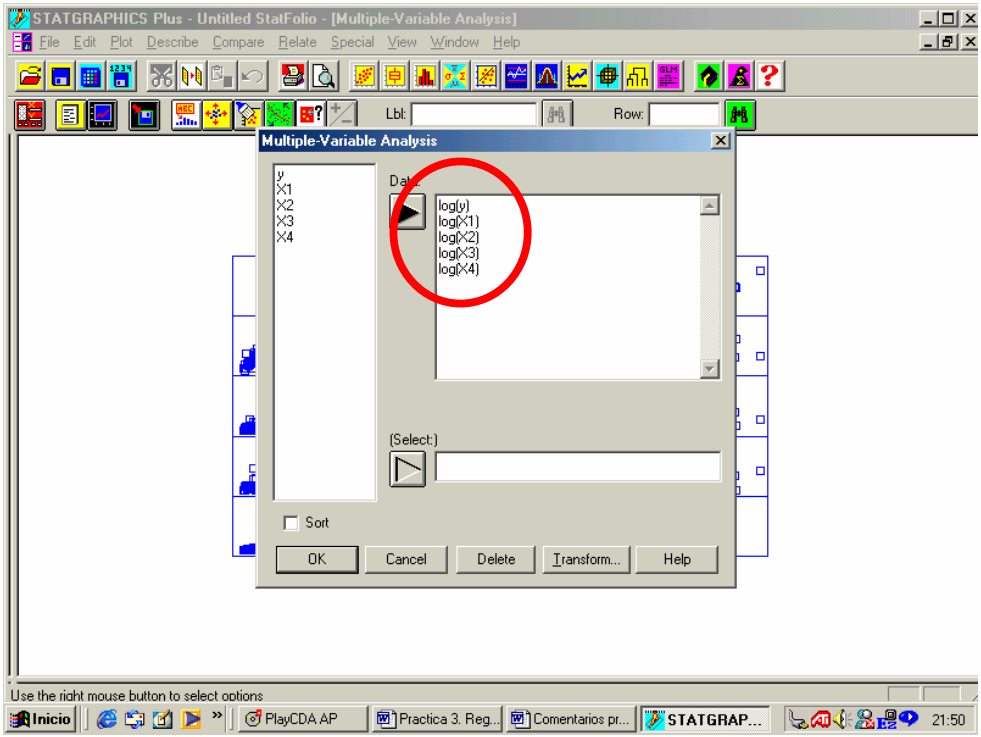
Se obtiene el siguiente Menú:



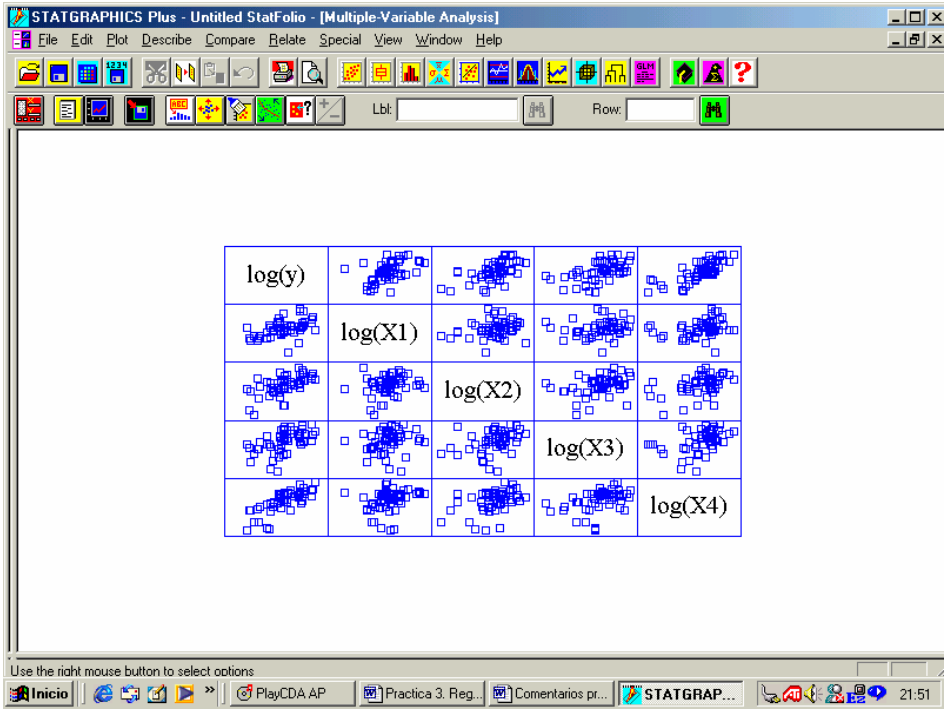
Obtendremos:



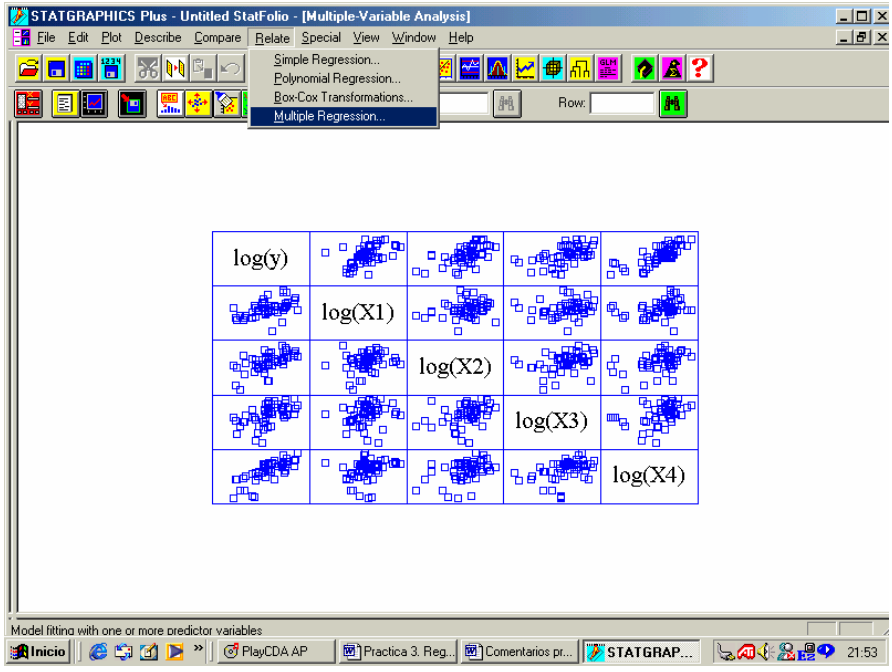
Como la relación entre los datos no es homocedástica, tenemos que transformar los datos tomando logaritmos.

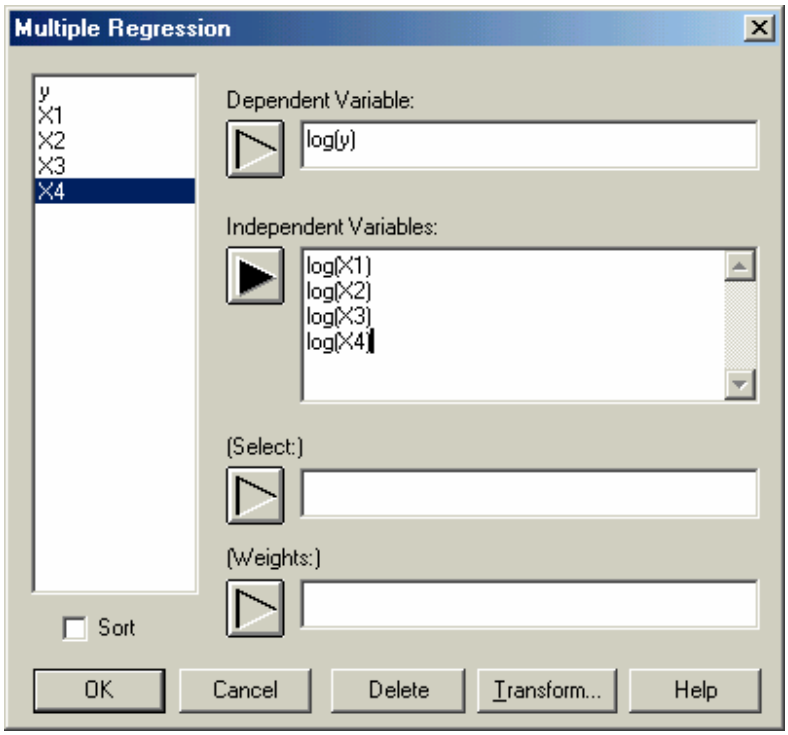


Ahora se obtiene:



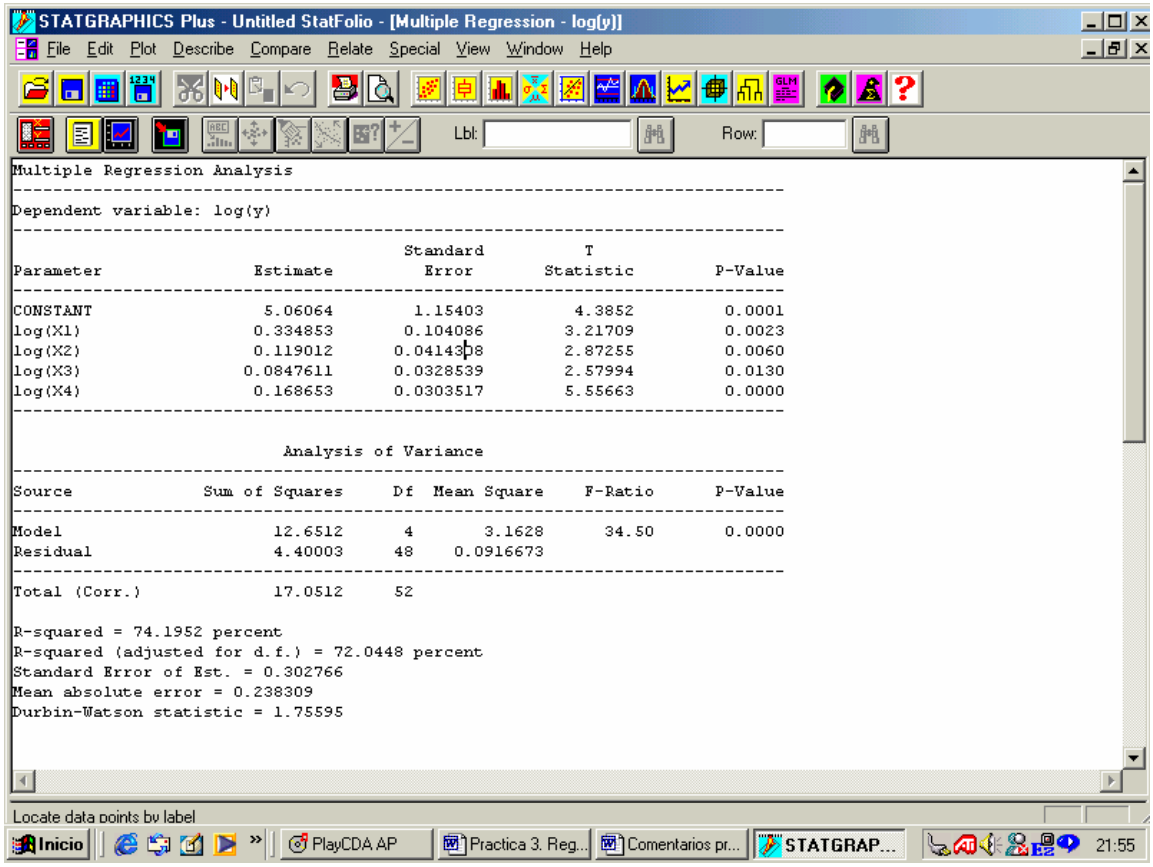
Los datos transformados tienen relación lineal, por tanto podemos estimar un modelo entre la variable respuesta Y, y las variables explicativas X´s





No hay que olvidar que si hemos hecho una transformación con los datos hay que mantenerla a lo largo de todo nuestro análisis

El resultado de la estimación es análogo a la regresión simple, pero ahora incluyendo más variables:



3. **Interpretación del modelo.**

Una vez que se ha realizado la estimación del modelo de regresión, deben interpretarse sus resultados atendiendo a los siguientes aspectos:

1. Observar los **valores de los estadísticos t** (o sus correspondientes valores p) para comprobar qué variables son o no significativas. Como regla práctica se puede decir que ***toda variable que tenga un estadístico t mayor de 2 (o un valor p menor que 0.05) es significativa.*** Si una variable no es significativa no debe incluirse en el modelo. En nuestro caso, todas las variables del modelo son significativas ya que sus t son mayores que 2.

2. **Signo de los parámetros estimados**. Proporcionan información sobre la relación entre cada una de las variables explicativas y la variable respuesta. Así cuando el signo del estimador sea positivo, nos indicará que al crecer la variable explicativa X_i (manteniéndose las demás constantes) también lo hará la variable respuesta Y. Si el signo es negativo, al aumentar la variable explicativa

(manteniéndose todas las demás constantes) la variables repuesta decrecerá. En nuestro caso, todos los signos de los estimadores son positivos, lo que implica que si cualquiera de las variables X´s aumenta su valor, también aumentará el valor de Y.

3. **Valores de los parámetros estimados**: Proporcionan información sobre cómo se transmite un incremento de la variable explicativa a la variable respuesta. Si aumentamos una de las X's, por ejemplo X_i , el incremento se transmitirá a la variable respuesta multiplicado por el valor del parámetro, es decir:

$$\Delta X_i \Rightarrow \Delta Y = \hat{\beta}_i * \Delta X_i$$

Donde $\hat{\beta}_i$ es el valor numérico del parámetro estimado con el modelo de regresión.

Es importante tener presente que si se ha realizado una transformación logarítmica de los datos, no se trataría de incrementos en las variables, sino de tasas de incremento. Por tanto diríamos que si la variable X_i se incrementa en un determinado porcentaje, la respuesta (supuestas las demás variables explicativas constantes) lo haría en el mismo porcentaje multiplicado por $\hat{\beta}_i$.

En nuestro casi, por tratarse de un modelo logarítmico, podremos decir, a modo de ejemplo que cuando la variable X1 aumenta un 1%, la variable Y lo hará en 1% x 0,33 = 0,33%.

Si no hubiera logaritmos diríamos que si X se incrementa en 1 unidad Y se incrementaría en 0.33x1=0.33 unidades.

4. **Coeficiente de determinación:** como en el caso de regresión simple, el coeficiente de determinación mide qué porcentaje de la variable respuesta Y es explicada por los regresores X´s. En el caso de la regresión múltiple, el coeficiente de determinación presenta un problema: el valor del coeficiente aumenta al añadir variables al modelo (sean significativas o no). Por evitar este problema se define un nuevo coeficiente, el *coeficiente de determinación corregido* (o R Squared adjusted como se denomina en Statgraphics) que *es el que debe considerarse siempre en regresión múltiple*. En el ejemplo analizado, con R2 corregido es 72.04, esto significa que con el modelo estimado se explica el 72% de la variabilidad de la variable respuesta Y

```
R-squared = 74.1952 percent
R-squared (adjusted for d.f.) = 72.0448 percent
Standard Error of Est. = 0.302766
Mean absolute error = 0.238309
Durbin-Watson statistic = 1.75595
```

5. **Intervalos de confianza:** como en regresión simple, el intervalo de confianza para el valor de un parámetro puede calcularse mediante la formula:

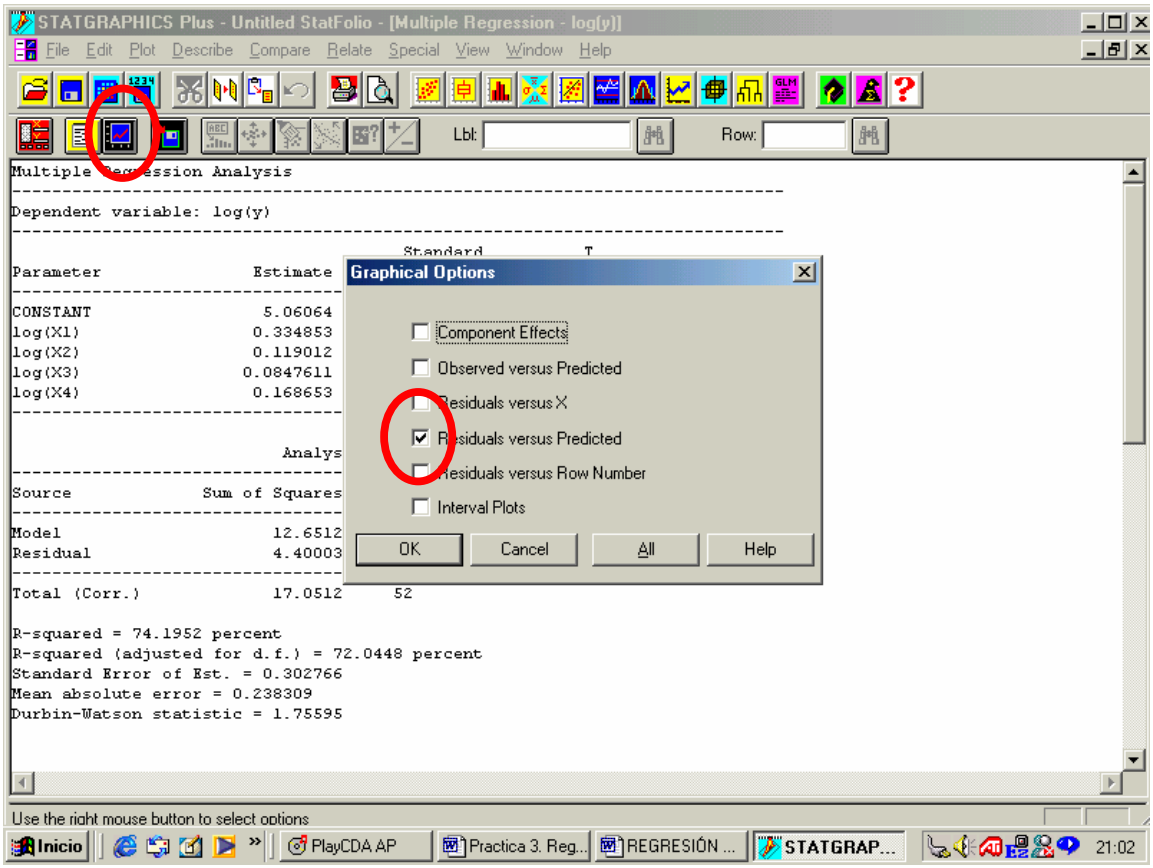
$$IC = \beta_i \pm 2\ SE(\beta_i)$$

Por tanto, para el valor β_1 correspondiente a la variable X1 tendremos que el intervalo de confianza será:

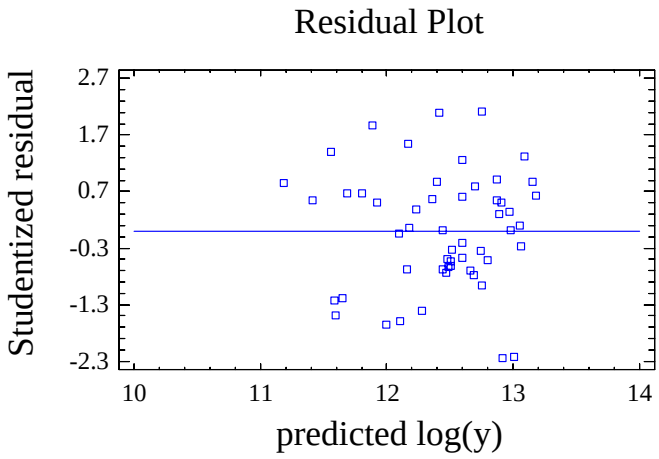
$$0{,}33 \pm 2 \times 0{,}10 = (0{,}55;\ 0{,}13)$$

4. Diagnosis y predicciones del modelo

Cuando un modelo tiene sus parámetros significativos, aun no podemos utilizarlo en la previsión. Debe realizarse una última comprobación conocida como diagnosis del modelo. Esta diagnosis se centrará fundamentalmente en comprobar que los residuos del modelo no tienen estructura.

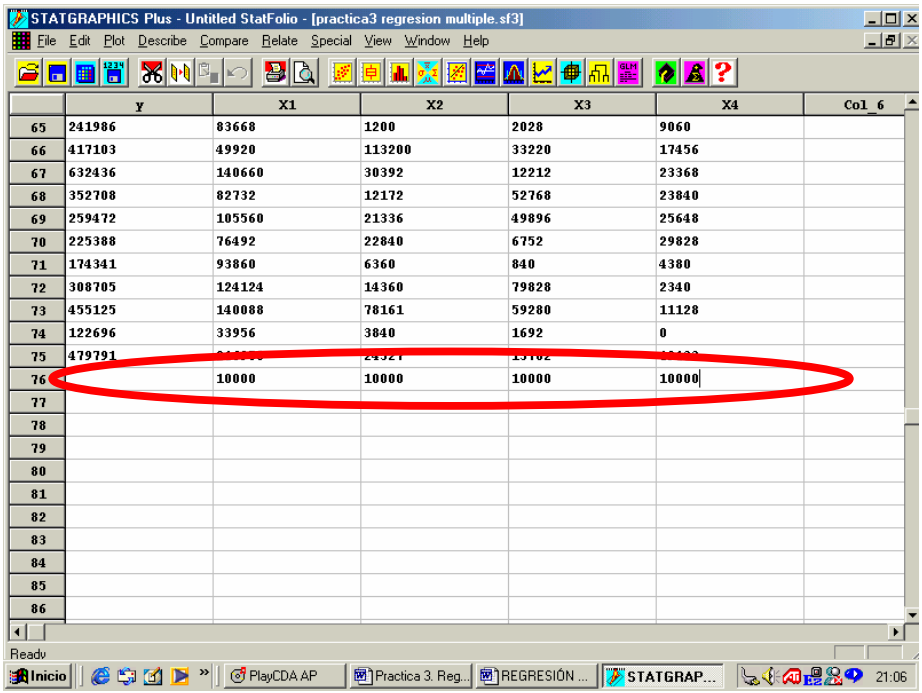


Obtenemos:



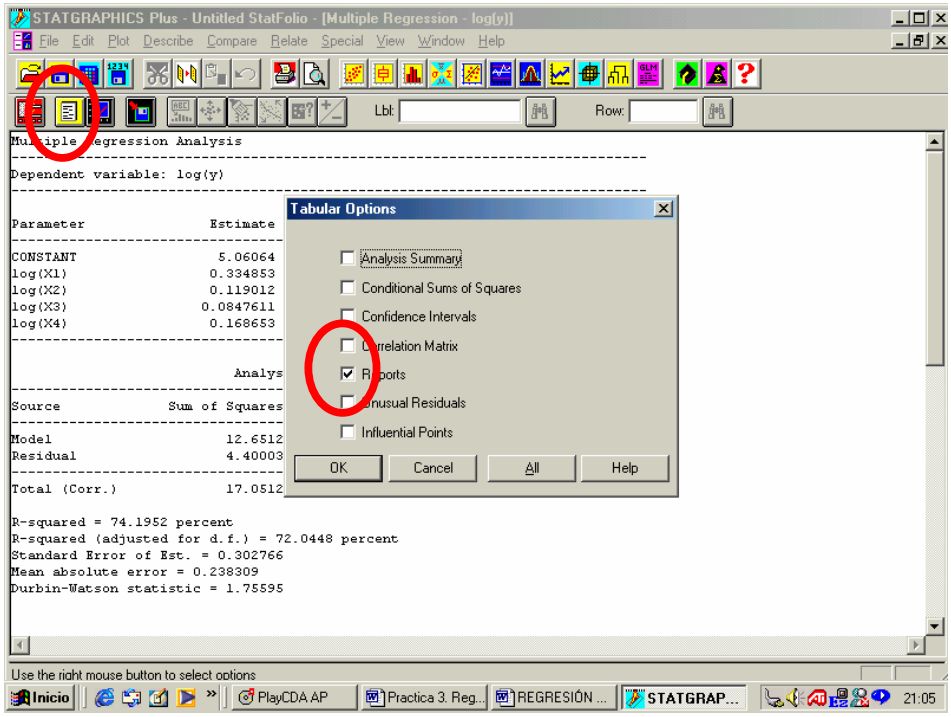
Que podemos aceptar como residuos carentes de estructura.

Comprobada la diagnosis del modelo, podemos utilizarlo para realizar previsiones:
Para ello, previamente en la hoja de datos debemos introducir (en la primera fila libre) los valores de los regresores X´s para los cuales queremos predecir el valor de Y

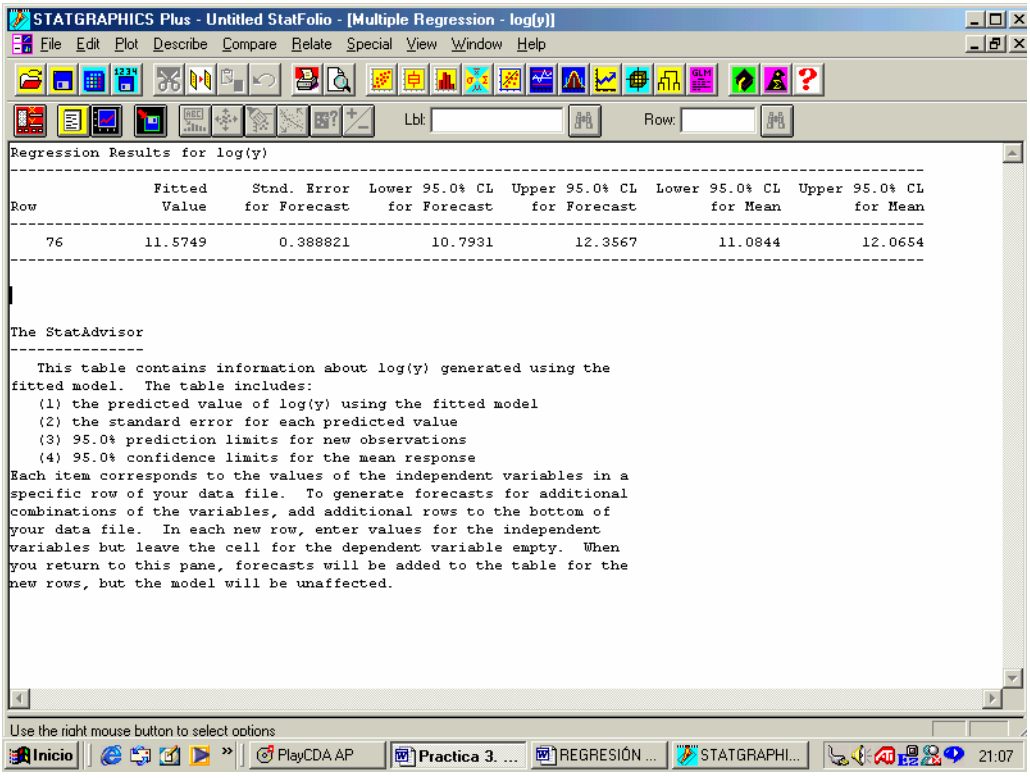


	y	X1	X2	X3	X4	Col_6
65	241986	83668	1200	2028	9060	
66	417103	49920	113200	33220	17456	
67	632436	140660	30392	12212	23368	
68	352708	82732	12172	52768	23840	
69	259472	105560	21336	49896	25648	
70	225388	76492	22840	6752	29828	
71	174341	93860	6360	840	4380	
72	308705	124124	14360	79828	2340	
73	455125	140088	78161	59280	11128	
74	122696	33956	3840	1692	0	
75	479791					
76		10000	10000	10000	10000	
77						
78						
79						
80						
81						
82						
83						
84						
85						
86						

A continuación haremos:



Obteniendo:



Vemos como Statgraphics nos muestra el valor predicho por nuestro modelo de regresión (*Fited value= 11.5749)* y además nos muestra el valor del Error Estándar cometido en la previsión (0.38821) a partir del cual calcula dos intervalos de confianza:

- el primero de ellos para la predicción del valor puntual de la variable respuesta cuando las variables explicativas toman el valor seleccionado. (10.79,12.35)

- el segundo nos da el intervalo de confianza para el valor medio de la variable repuesta en el caso de que tuviésemos muchas observaciones en las que las variables explicativas tomasen el valor seleccionado. (11.08, 12.06)

Obsérvese como la longitud del intervalo para la previsión del valor medio de una observación es menor que la longitud del intervalo de confianza cuando se predice el valor puntual de la nueva observación. Es este un resultado razonable pues resulta más sencilla la previsión del valor medio, y el modelo de regresión nos da este valor con una mayor precisión (intervalo de confianza más pequeño).

Ejercicios

Con los datos del fichero, calcular:

- Escribir la ecuación de regresión
- Construir un intervalo de confianza para los coeficientes de las variables X1 y X2
- ¿Son significativas estas variables?
- Cuanto se incrementa la variable Y cuando cada una de estas dos variables (manteniendo constantes todas las demás) se incrementa en un 10%
- Calcular el valor de Y cuando cada uno de los regresores X´s toma un valor de 15.000.