

Primera práctica de REGRESIÓN.

DATOS: fichero “practica regresión 1.sf3”

1. Objetivo:

El objetivo de esta práctica es aprender cuándo se puede utilizar el análisis de regresión.

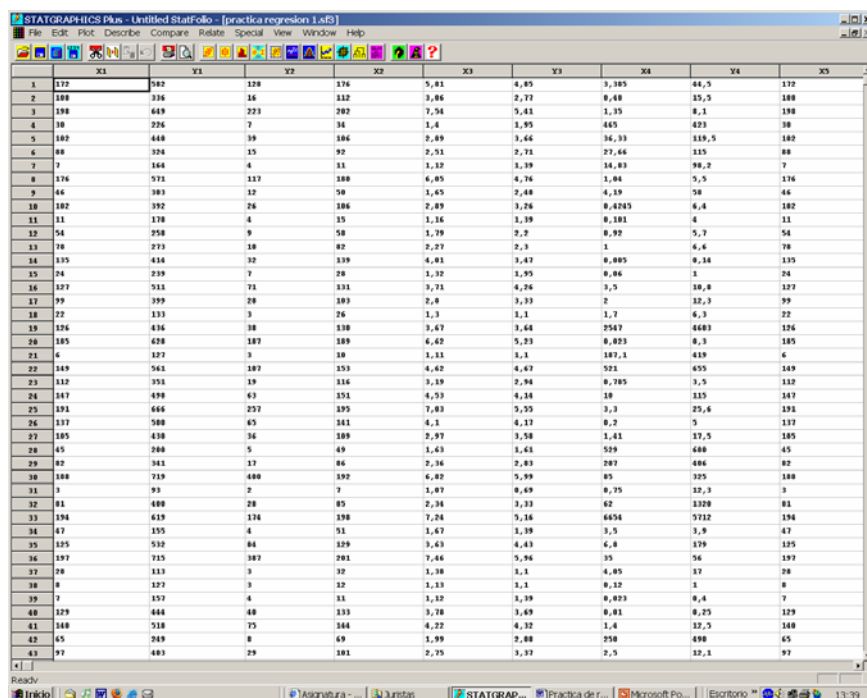
En la primera parte se presentan unos conjuntos de datos para que decidir si son adecuados o no. Esta decisión ha de tomarse teniendo en cuenta si los datos cumplen las hipótesis del modelo (Linealidad, homocedasticidad...)

En caso de que no las cumplan será preciso transformar los datos hasta conseguir que cumplan las hipótesis. Este proceso se desarrolla en el apartado 2 de la práctica.

Finalmente en el tercer apartado, se van a ajustar algunas regresiones simples. El objetivo es aprender a hacer el ajuste con datos adecuados y leer la ecuación de regresión. La interpretación y contrastes será el objetivo de la segunda práctica.

2. Análisis gráfico y transformaciones.

Vamos a utilizar el fichero “practica regresión 1.sf3”. El fichero contiene 6 pares de variables X-Y. El aspecto del fichero es:



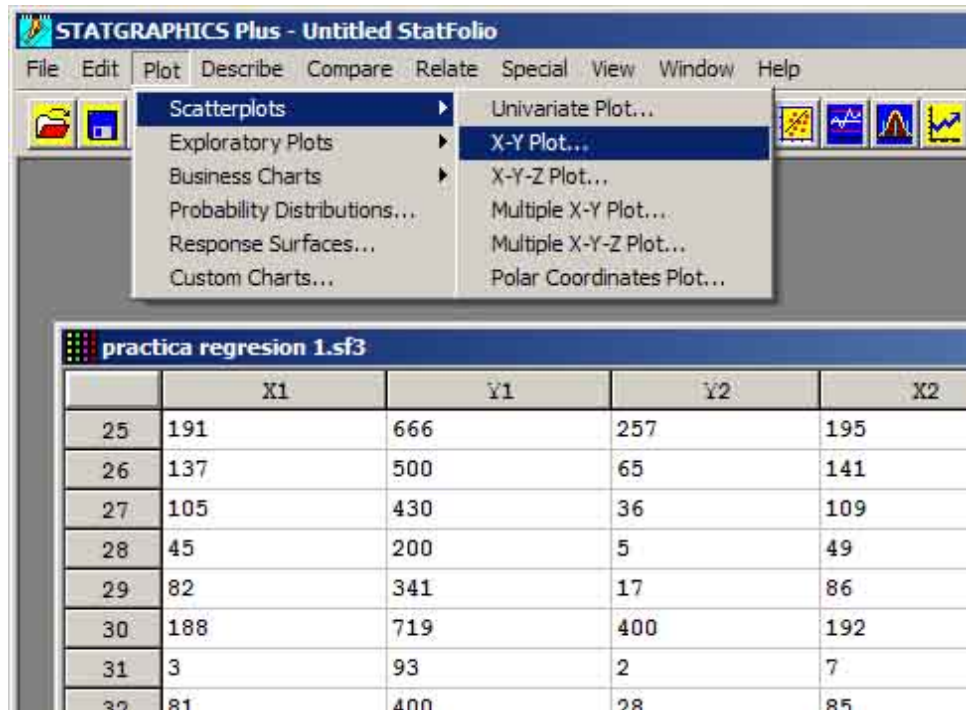
	X1	Y1	X2	Y2	X3	Y3	X4	Y4	X5
1	572	582	128	176	5,81	4,85	3,185	44,5	132
2	188	116	15	112	3,86	2,71	8,18	15,3	188
3	198	449	223	282	7,58	5,43	1,35	8,1	198
4	38	226	7	36	1,4	1,85	465	423	38
5	182	448	39	186	2,89	1,44	16,33	119,5	182
6	88	124	15	92	2,51	2,71	27,44	115	88
7	7	164	4	11	1,12	1,39	14,83	98,2	7
8	176	571	113	188	4,85	4,76	1,84	5,5	176
9	46	183	12	58	1,45	2,48	4,19	58	46
10	182	192	26	186	2,89	1,24	8,4243	6,4	182
11	11	178	4	15	1,16	1,39	8,181	4	11
12	54	258	9	58	1,79	2,2	8,92	5,7	54
13	78	273	18	82	2,27	2,3	1	4,6	78
14	135	414	32	139	4,81	1,47	8,885	8,14	135
15	24	239	7	28	1,32	1,95	8,86	1	24
16	127	511	71	131	3,71	4,26	3,5	18,8	127
17	79	199	28	183	2,8	3,33	2	12,3	79
18	22	133	3	26	1,3	1,1	1,7	6,3	22
19	126	416	18	118	3,47	1,44	2547	4683	126
20	185	628	187	189	6,42	5,23	8,823	8,3	185
21	4	127	3	18	1,11	1,1	187,1	419	4
22	149	561	182	153	4,42	4,47	121	655	149
23	112	351	19	116	3,19	2,84	8,785	3,5	112
24	167	498	63	151	4,53	4,14	18	115	167
25	191	666	252	195	7,83	5,55	3,3	25,6	191
26	117	588	85	161	4,1	4,17	8,2	5	117
27	185	418	16	189	2,97	1,58	1,41	17,5	185
28	45	288	5	49	1,43	1,41	109	488	45
29	82	341	17	86	2,14	2,83	287	486	82
30	188	719	488	192	6,82	5,99	85	125	188
31	3	93	2	7	1,87	8,49	8,75	12,3	3
32	81	488	28	85	2,36	3,33	62	1328	81
33	194	619	126	198	7,24	5,16	4854	7312	194
34	47	155	4	51	1,47	1,39	3,5	1,9	47
35	125	512	84	129	3,43	4,43	6,8	129	125
36	197	715	183	281	7,46	5,84	35	56	197
37	28	113	3	32	1,38	1,1	4,85	17	28
38	8	123	3	12	1,13	1,1	8,12	1	8
39	7	157	4	11	1,12	1,39	8,823	8,4	7
40	129	444	48	133	3,78	1,69	8,81	8,23	129
41	148	518	75	144	4,22	4,32	1,4	12,3	148
42	45	248	8	49	1,99	2,88	258	488	45
43	97	483	29	181	2,75	3,37	2,5	12,1	97

El primer paso en cualquier regresión es comprobar las hipótesis del modelo:

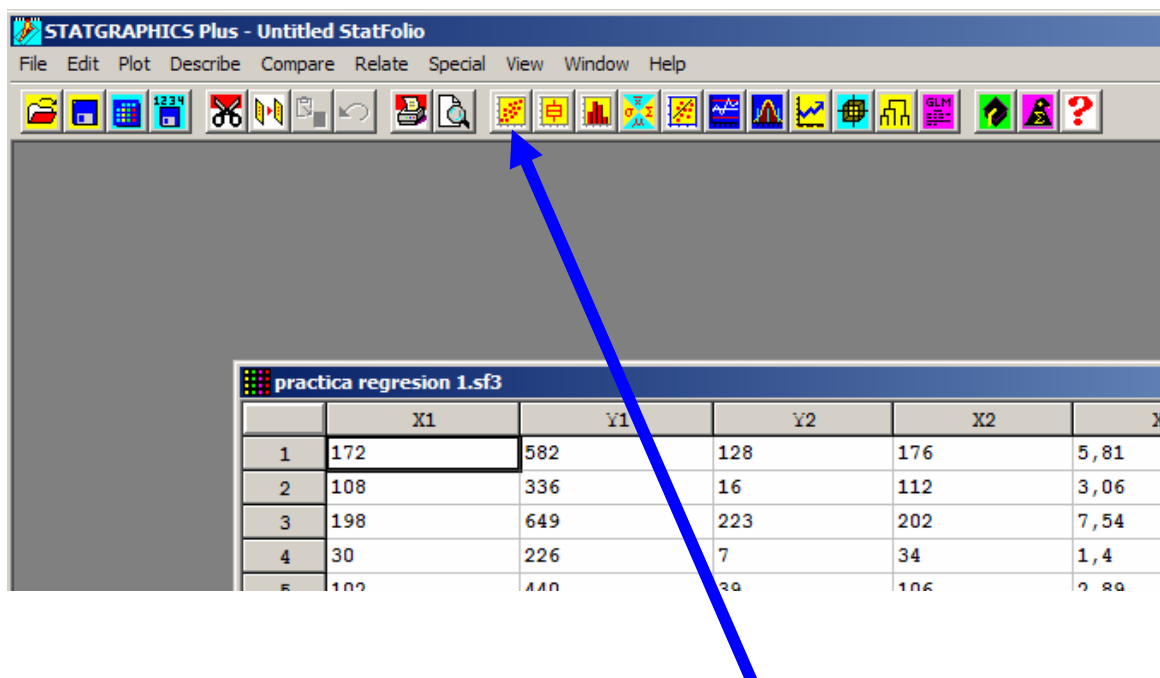
a. *Linealidad.*

Los datos deben ser lineales, lo que implica que su gráfico de dispersión XY debe presentar un aspecto razonablemente recto.

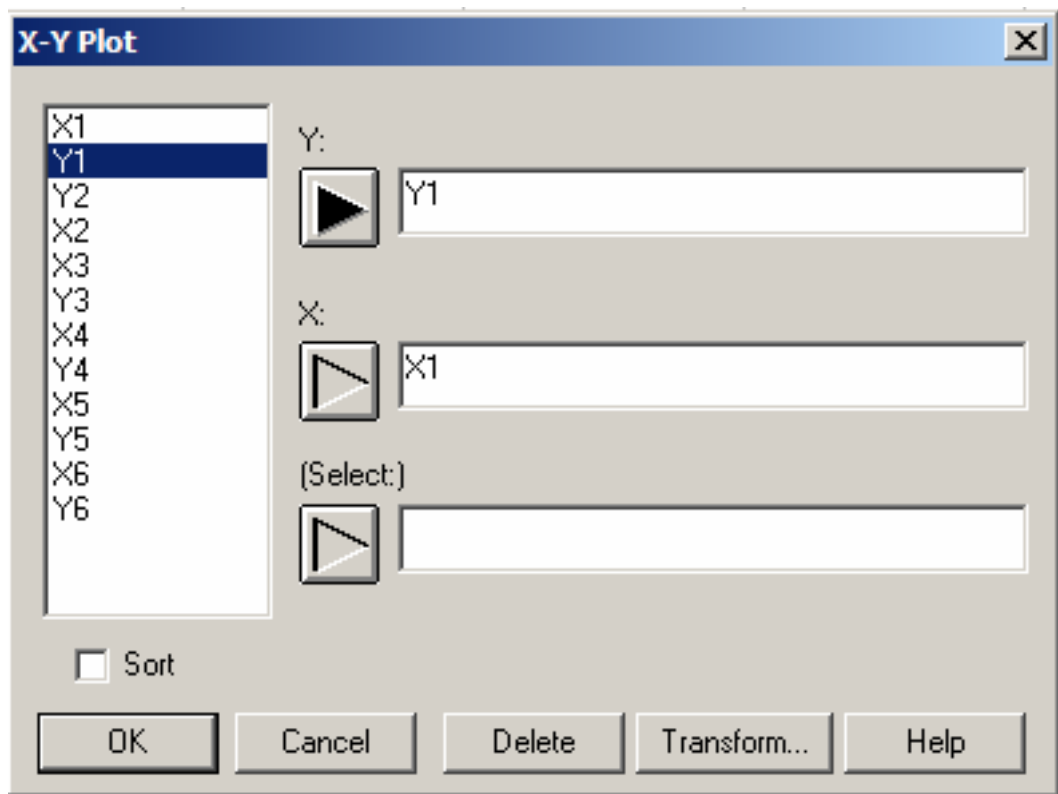
Para realizar el gráfico se pincha en PLOT-SCATTERPLOTS-XY PLOT:



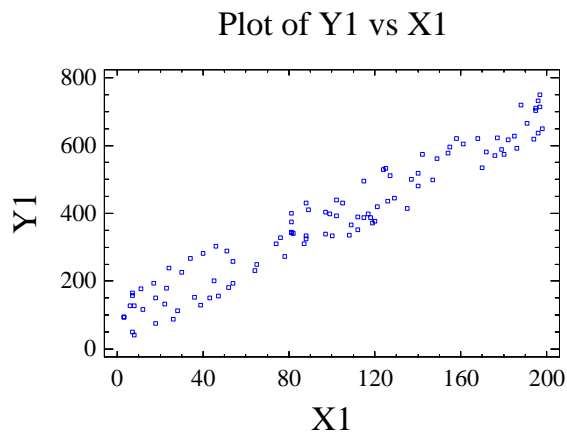
También puede hacerse directamente pinchando en el icono SCATTERPLOT:



Se abre el menú del gráfico y seleccionamos las variables adecuadas:



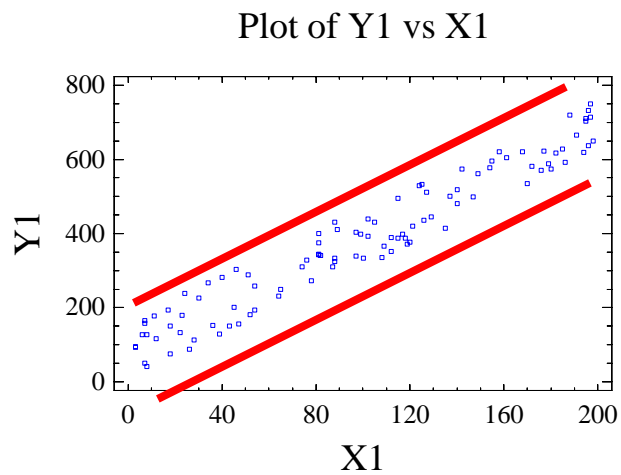
Si hacemos un gráfico de dispersión con los datos de las variables X1-Y1 obtenemos:



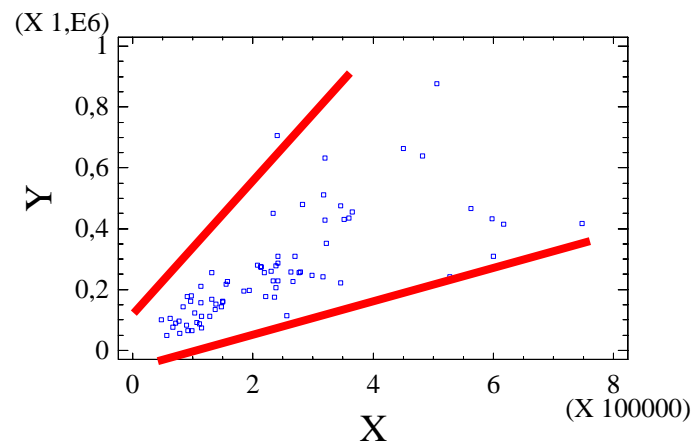
Este gráfico presenta un aspecto claramente lineal por lo que los datos X1, Y1 cumplen la primera hipótesis.

b. Hipótesis de homocedasticidad:

Los datos deben tener la misma varianza (grosor) por lo que el aspecto en el gráfico de dispersión debe ser de grosor constante. En el caso que estamos estudiando,



vemos que el grosor de los datos es constante. Por lo que se cumple la segunda hipótesis. Si los datos hubieran sido:



la dispersión (grosor) no sería constante y no se cumpliría la segunda hipótesis. Estos datos requieren una transformación.

c. Hipótesis de independencia.

Esta hipótesis es muy importante. Simplemente hay que tener en cuenta que la regresión no es adecuada para ser utilizada con datos temporales. Por ejemplo, no se debe aplicar la regresión a datos como los siguientes:

Año	Coches vendidos (Miles)	Teléfonos vendidos (Miles)
1981	768	124
1982	780	135
1983	793	236
1984	801	397
1984	799	450
1986	803	457

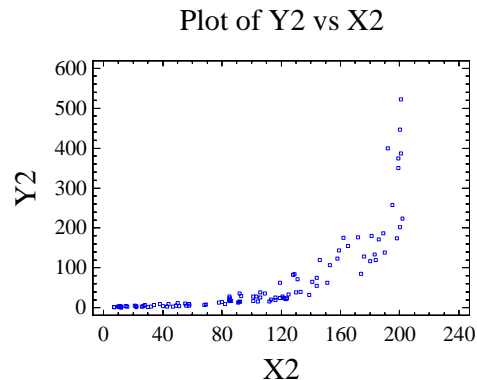
Los datos temporales son aquellos en cada observación se refiere a un periodo de tiempo concreto.

d. Hipótesis de Normalidad

Esta hipótesis se comprobará en la diagnosis del modelo (en la segunda práctica)

Transformaciones:

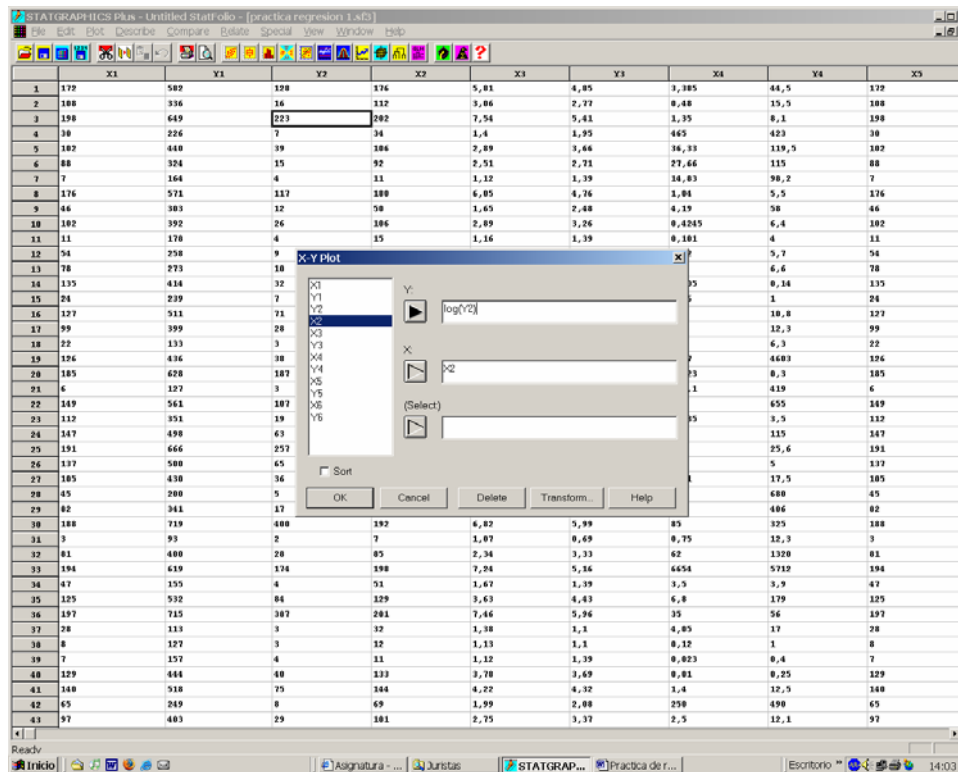
Si los datos no cumplen las hipótesis, es preciso transformarlos. Vamos a hacer un gráfico de las variables X2-Y2 del fichero de prácticas:



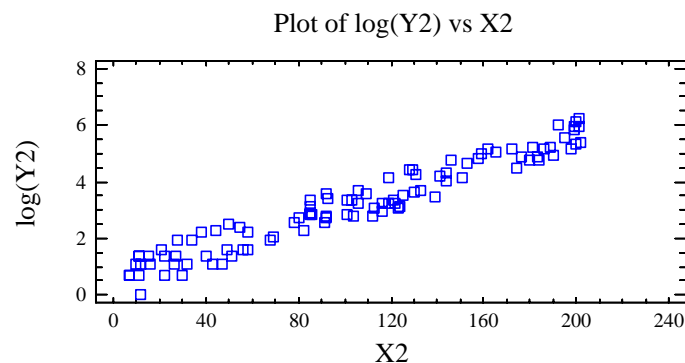
Evidentemente estos datos no son lineales. Es preciso transformarlos. Para ver qué transformación es necesaria, en el menú del gráfico de dispersión se van probando diversas transformaciones. Las más habituales son:

- Logaritmos: $\text{Log}(X1)$ o $\text{Log}(Y1)$
- Logaritmo de ambas variables.
- Exponenciales: $\exp(X1)$ o $\exp(Y1)$
- Inversas: $1/Y1$ o $1/X1$
- Cuadrados: $(X1)^2$ o $(Y1)^2$
- Raíz cuadrada: $\text{SQRT}(X1)$ o $\text{SQRT}(Y1)$

En nuestro caso la transformación adecuada es $\text{Log}(Y2)$ frente $X2$ sin transformar. En el menú del gráfico de dispersión se escribe en el lugar de la variable Y “ $\log(Y2)$ ”. En el lugar de la variable X se escribe “ $X2$ ”.



Obteniendo un gráfico lineal y homocedástico.



En este caso, habrá que ajustar el modelo de regresión:

$$\log(y_i) = \beta_0 + \beta_1 x_i + u_i$$

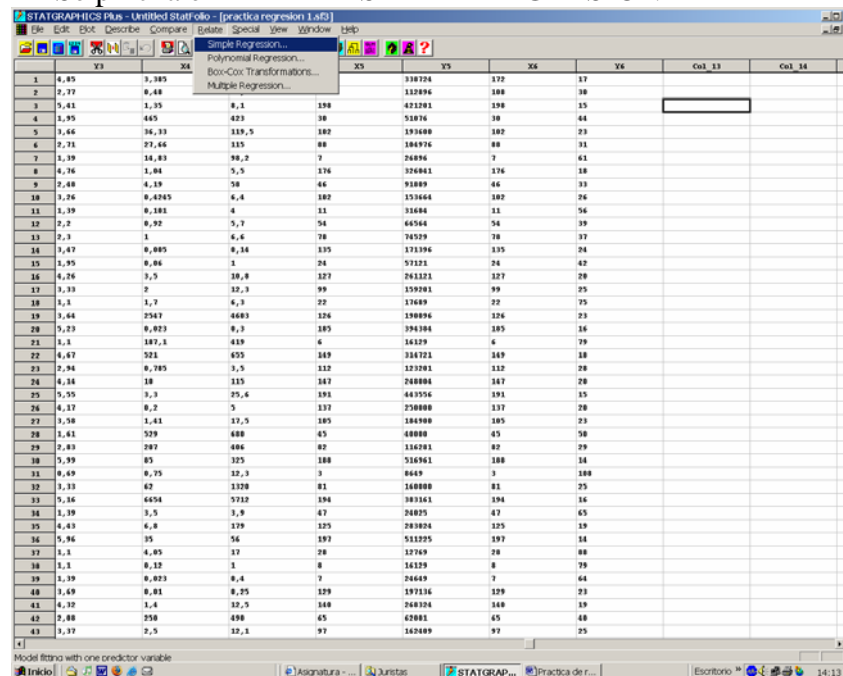
Primera parte de la práctica: Estudiar las parejas de variables X1-Y1, X2-Y2, etc y determinar si cumplen las hipótesis de linealidad y homocedasticidad para realizar el análisis de regresión. En caso de que no las cumplan, transformar las variables hasta obtener una relación lineal y homocedástica.

Variable	Variable	Trasformación de X	Transformación de Y
X1	Y1	No es necesario	No es necesario
X2	Y2	No es necesario	Log(Y2)
X3	Y3		
X4	Y4		
X5	Y5		
X6	Y6		

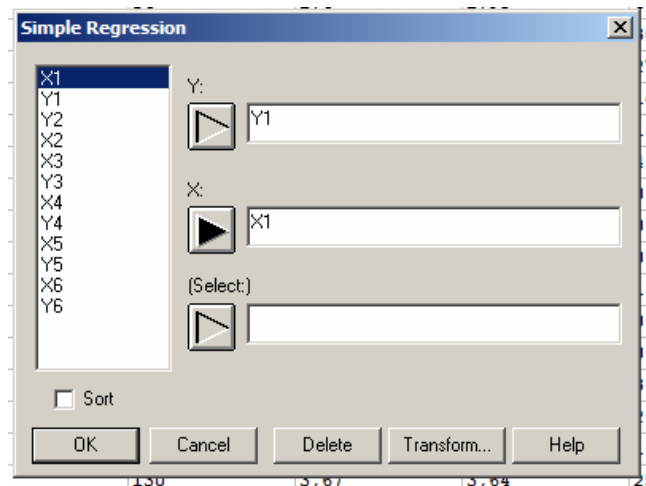
3. Regresión simple. Ajuste

Una vez que los datos son normales y homocedásticos hay que ajustar un modelo de regresión. El proceso es muy simple:

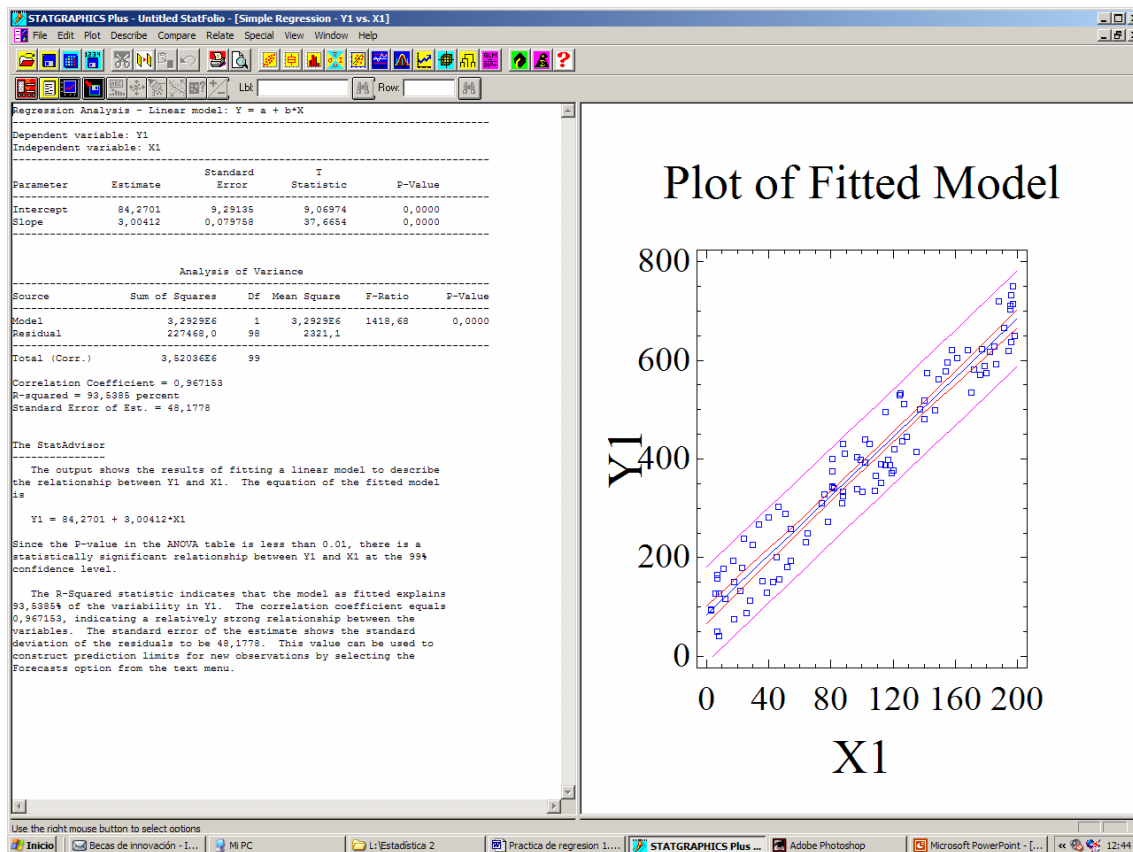
- Se pincha en RELATE-SIMPLE REGRESIÓN



- Se obtiene un menú que pide la variable Y y la variable X:



El resultado es:



La ventana de la izquierda presenta la regresión. Pinchando dos veces sobre la ventana se amplía.

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: Y1

Independent variable: X1

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	84,2701	9,29135	9,06974	0,0000
Slope	3,00412	0,079758	37,6654	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	3,2929E6	1	3,2929E6	1418,68	0,0000
Residual	227468,0	98	2321,1		
Total (Corr.)	3,52036E6	99			

Correlation Coefficient = 0,967153

R-squared = 93,5385 percent

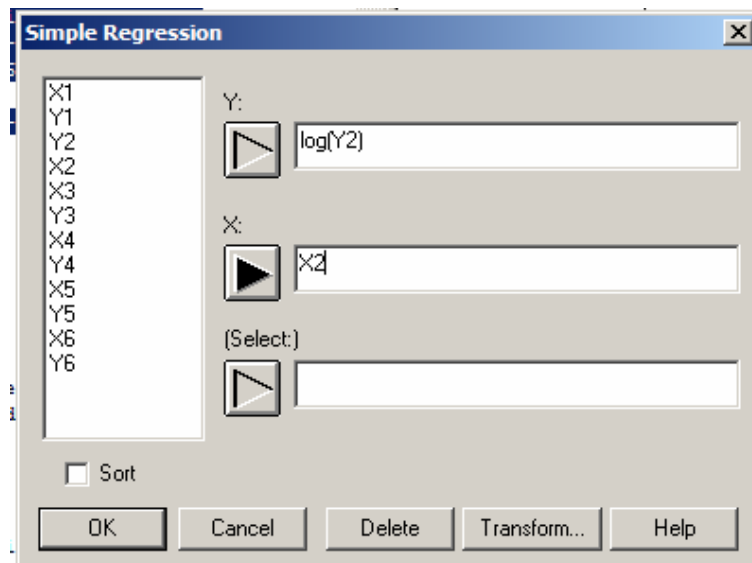
Standard Error of Est. = 48,1778

La ecuación de regresión se obtiene en la columna Estimate. El valor Intercept es la estimación de la constante y Slope el valor de la pendiente. En este caso la regresión vale:

$$\hat{y}_i = 84.27 + 3x_i$$

\hat{y}_i indica valor previsto de Y para x_i .

Si hubiéramos realizado la regresión con los datos X2-Y2, en los que era preciso transformar Y2 a Log(Y2), tendríamos que haber puesto en el menú de regresión como variable Y, la transformación Log(Y2).



Y el resultado es:

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: $\log(Y2)$

Independent variable: $X2$

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	0,608308	0,0797732	7,62547	0,0000
Slope	0,0249962	0,000665152	37,5798	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	227,979	1	227,979	1412,24	0,0000
Residual	15,8202	98	0,161431		
Total (Corr.)	243,799	99			

Correlation Coefficient = 0,967011

R-squared = 93,511 percent

Standard Error of Est. = 0,401784

$$\log(\hat{y}_{2i}) = 0,61 + 0.025x_{2i}$$

Segunda parte de la práctica: Estimar las regresiones simples para las parejas de variables $X1-Y1$, $X2-Y2$, etc teniendo en cuenta las transformaciones adecuadas que se encontraron en la primera parte de la práctica.

Variable	Variable	Regresión:
X1	Y1	$\hat{y}_i = 84.27 + 3x_i$
X2	Y2	$\log(\hat{y}_{2i}) = 0,61 + 0.025x_{2i}$
X3	Y3	
X4	Y4	
X5	Y5	
X6	Y6	