

# Quinta práctica de REGRESIÓN.

DATOS: fichero “practica regresión 5.sf3”

## 1. Objetivo:

Cuando la población que se analiza puede dividirse en grupos según una cualidad, es necesario modelizar la pertenencia al grupo.

El propósito de esta práctica es familiarizarse con problemas de este tipo, para lo cual es necesario conseguir los siguientes objetivos:

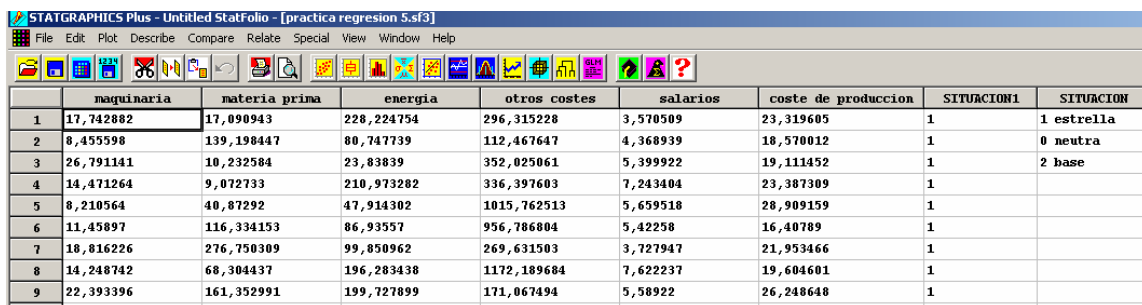
- Construir modelos de regresión con presencia de variables cualitativas que reflejen la división en grupos de la población estudiada.
- Identificar interacciones entre las variables explicativas y las variables cualitativas, incluyéndolas en el modelo
- Interpretar adecuadamente los modelos de regresión con variables cualitativas, extrayendo conclusiones sobre la estructura y comportamiento del fenómeno analizado.

## Conocimientos necesarios de otras prácticas

*Estimación, interpretación de parámetros y diagnosis de modelos de regresión múltiple*

## 2. Generación de variables dicotómicas o cualitativas.

El archivo *practica regresion5.sf3* contiene datos sobre las factorías de una empresa multinacional especializada en la producción acero. El aspecto del archivo es:

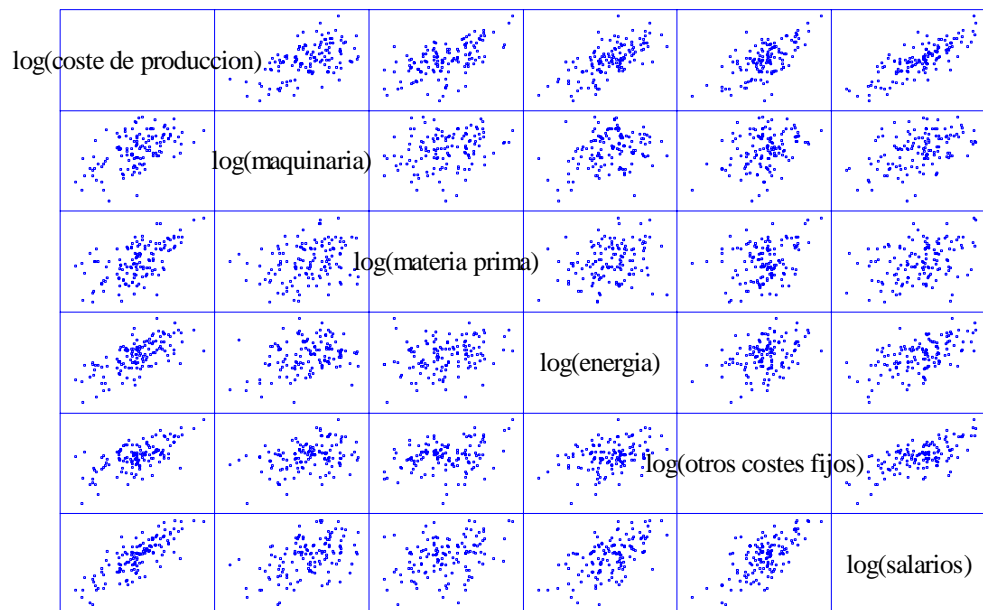


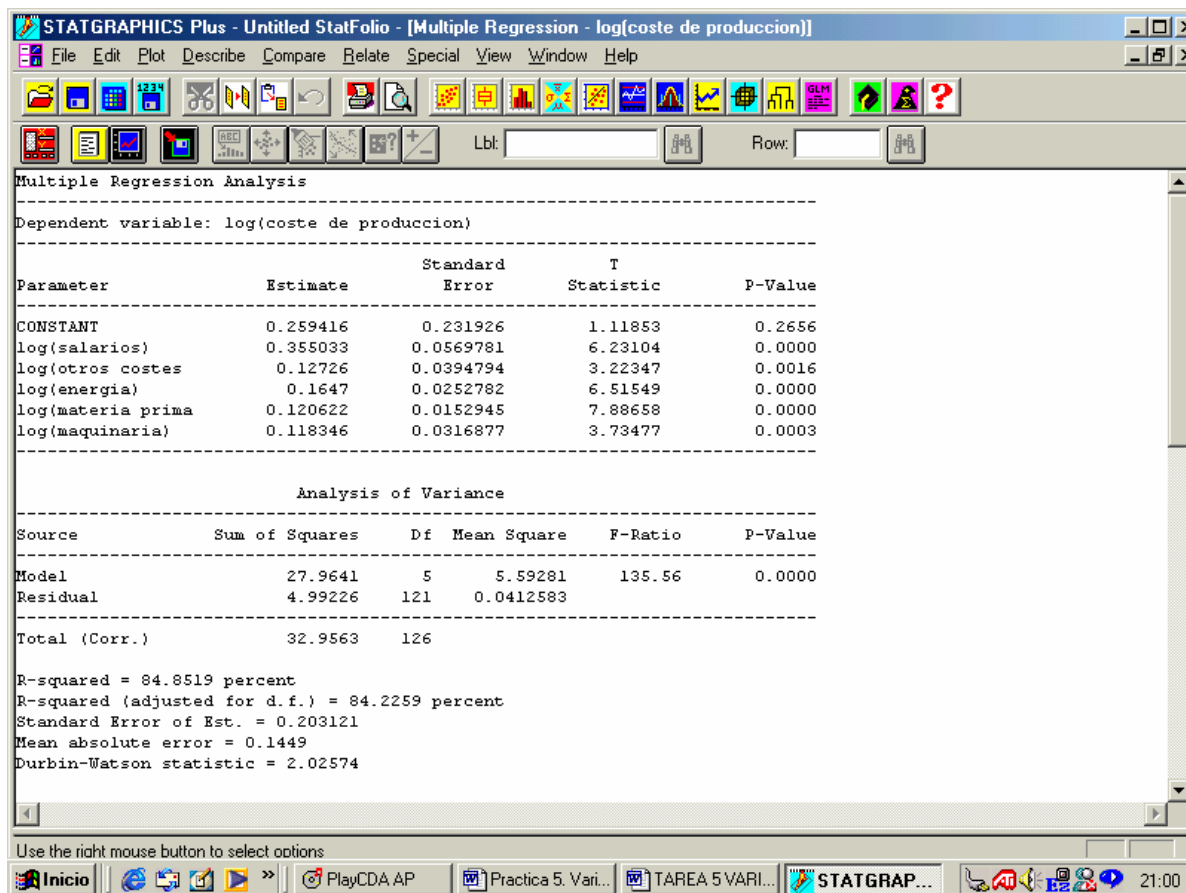
	maquinaria	materia prima	energia	otros costes	salarios	coste de produccion	SITUACION1	SITUACION
1	17,742882	17,090943	228,224754	296,315228	3,570509	23,319605	1	1 estrella
2	8,455598	139,198447	80,747739	112,467647	4,368939	18,570012	1	0 neutra
3	26,791141	10,232584	23,83839	352,025061	5,399922	19,111452	1	2 base
4	14,471264	9,072733	210,973282	336,397603	7,243404	23,387309	1	
5	8,210564	40,87292	47,914302	1015,762513	5,659518	28,909159	1	
6	11,45897	116,334153	86,93557	956,786804	5,42258	16,40789	1	
7	18,816226	276,750309	99,850962	269,631503	3,727947	21,953466	1	
8	14,248742	68,304437	196,283438	1172,189684	7,622237	19,604601	1	
9	22,393396	161,352991	199,727899	171,067494	5,58922	26,248648	1	

Las variables son:

- *coste de producción*: coste por unidad producida
- *salarios*: coste por hora trabajada
- *energías*: costes energéticos
- *materia prima*: coste de las materias primas
- *maquinaria*: coste de depreciación de la maquinaria utilizada en la producción.

Con estas variables, puede estimarse un modelo de regresión múltiple entre la variable coste de producción y el resto de las variables explicativas:





Obsérvese que el modelo está en logaritmos debido a la heterocedasticidad de los datos.

La empresa trata de situar sus factorías en emplazamientos preferenciales en función de la disponibilidad y coste de la materia prima utilizada en la producción.

Las factorías se dividen en tres grupos (*estrella*, *base*, *neutra*) dependiendo de su emplazamiento. Estos tres grupos aparecen en la variable *situación 1* del fichero de datos.

Para introducir en el modelo de regresión la pertenencia a estos grupos, deben generarse las variables dummy correspondientes. Para ello, se hará:

1.- Seleccionamos la primera columna vacía del fichero de datos:

STATGRAPHICS Plus - Untitled StatFolio - [practica5 regresion multiple.sf3]

	coste de produccion	SITUACION1	SITUACION	Col_9	Col_10	Col_11	Col_12
1	23.319605	1	1 estrella				
2	18.570012	1	0 neutra				
3	19.111452	1	2 base				
4	23.387309	1					
5	28.909159	1					
6	16.40789	1					
7	21.953466	1					
8	19.604601	1					
9	26.248648	1					
10	23.900756	1					
11	29.325299	1					
12	29.962872	1					
13	28.200567	1					
14	29.244735	1					
15	52.535003	0					
16	23.972342	1					
17	27.783748	1					
18	25.281348	1					
19	32.365672	1					
20	24.544759	1					
21	43.632965	1					
22	38.040672	1					

2.- Clickeamos con el botón derecho del ratón y activamos la opción “Generate data” del menú que nos aparece:

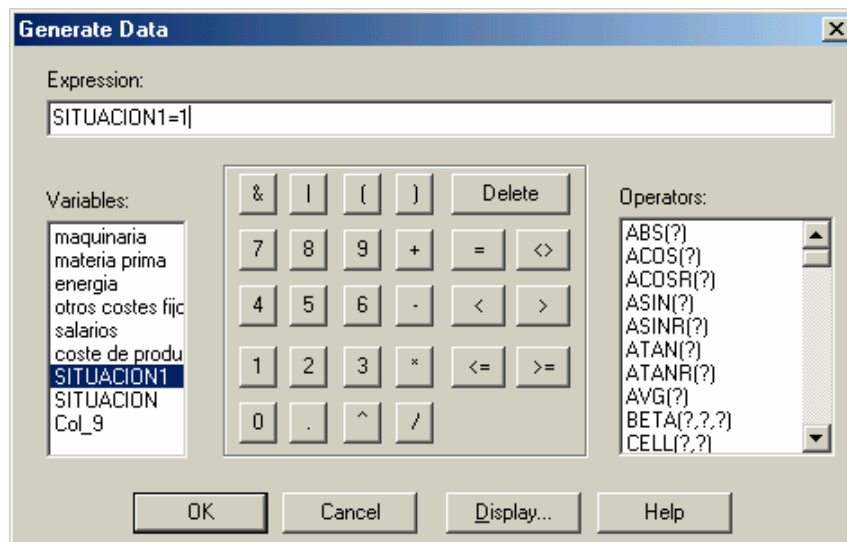
STATGRAPHICS Plus - Untitled StatFolio - [practica5 regresion multiple.sf3]

	coste de produccion	SITUACION1	SITUACION	Col_9	Col_10	Col_11	Col_12
1	23.319605	1	1 estrella				
2	18.570012	1	0 neutra				
3	19.111452	1	2 base				
4	23.387309	1					
5	28.909159	1					
6	16.40789	1					
7	21.953466	1					
8	19.604601	1					
9	26.248648	1					
10	23.900756	1					
11	29.325299	1					
12	29.962872	1					
13	28.200567	1					
14	29.244735	1					
15	52.535003	0					
16	23.972342	1					
17	27.783748	1					
18	25.281348	1					
19	32.365672	1					
20	24.544759	1					
21	43.632965	1					
22	38.040672	1					

Context Menu Options:

- Undo (Ctrl+Z)
- Cut (Ctrl+X)
- Copy (Ctrl+C)
- Paste (Ctrl+V)
- Paste Link (Ctrl+L)
- Insert
- Delete
- Modify Column... (Shift+F5)
- Generate Data... (Shift+F7)**
- Recode Data...
- Sort File...
- Update Formulas...
- Print... (F4)
- Print Preview... (Shift+F3)
- Save Data File (Shift+F12)
- Save Data File As... (F12)

Nos aparece un cuadro de dialogo donde debe definirse la condición de pertenencia al grupo analizado. Para las factorías estrella, situación1=1. Para las factorías base habrá que poner situacion1=2.



Al pulsar OK, nos aparece una variable dummy con 1 cuando la factoría está clasificada como estrella y 0 en caso contrario

	coste de produccion	SITUACION1	SITUACION	estrella	Col_10	Col_11	Col_12
1	23.319605	1	1 estrella	1			
2	18.570012	1	0 neutra	1			
3	19.111452	1	2 base	1			
4	23.387309	1		1			
5	28.909159	1		1			
6	16.40789	1		1			
7	21.953466	1		1			
8	19.604601	1		1			
9	26.248648	1		1			
10	23.900756	1		1			
11	29.325299	1		1			
12	29.962872	1		1			
13	28.200567	1		1			
14	29.244735	1		1			
15	52.535003	0		0			
16	23.972342	1		1			
17	27.783748	1		1			
18	25.281348	1		1			
19	32.365672	1		1			
20	24.544759	1		1			
21	43.632965	1		1			
22	38.040672	1		1			

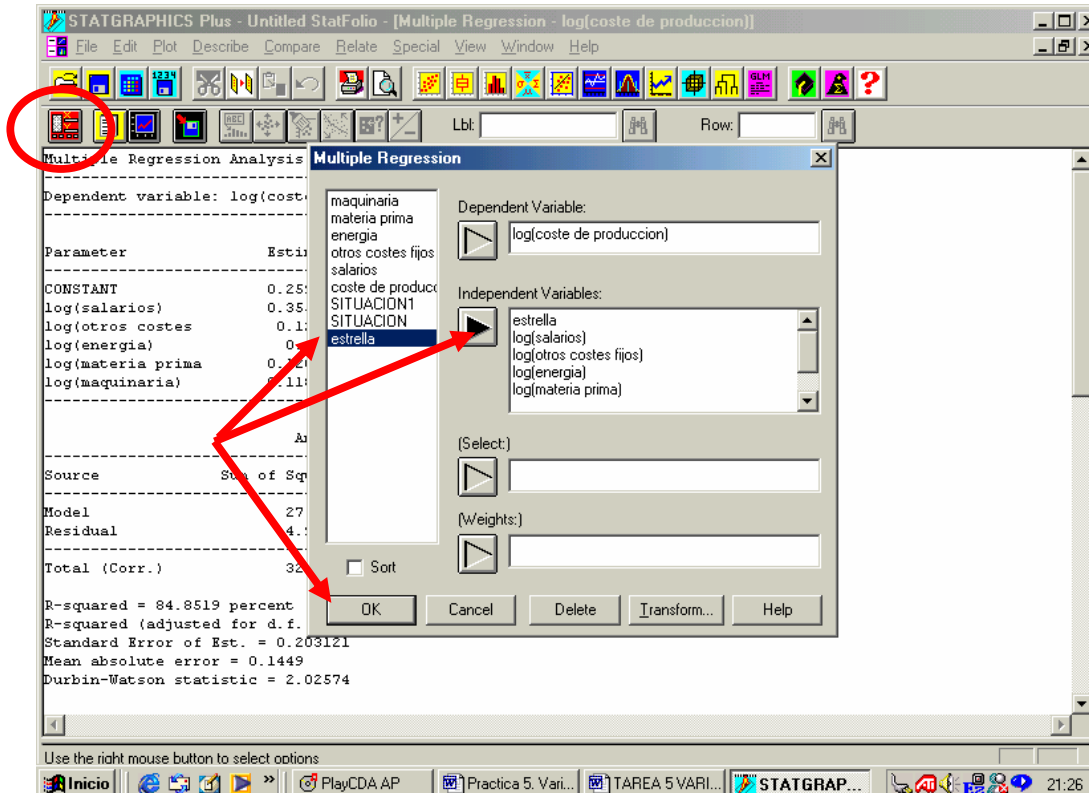
### Ejercicio:

Generar las variables dummie para las factorías “base” y “neutras”

### 3. Regresión con las variables dummies.

Las variables dummies deben introducirse en el modelo de forma análoga al resto de las variables, analizando si estas variables son o no significativas del mismo modo que cualquier otra variable explicativa.

Como ejemplo, puede comprobarse si las factorías denominadas *estrella* tienen unos costes de producción menores que el resto. Para ello introduciremos la variable dummie estrella dentro del modelo:



Obtendremos:

STATGRAPHICS Plus - Untitled StatFolio - [Multiple Regression - log(coste de produccion)]

File Edit Plot Describe Compare Relate Special View Window Help

Multiple Regression Analysis

Dependent variable: log(coste de produccion)

Parameter	Estimate	Standard Error	T Statistic	P-Value
Constant	0.975214	0.239167	4.07655	0.0001
estrella	-0.283898	0.0485562	-5.84678	0.0000
log(salarios)	0.26584	0.0527303	5.0415	0.0000
log(otros costes)	0.137163	0.0350147	3.91728	0.0001
log(energia)	0.122714	0.0235165	5.21819	0.0000
log(materia prima)	0.0936721	0.0143116	6.5452	0.0000
log(maquinaria)	0.0975135	0.0282965	3.44614	0.0008

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	29.0709	6	4.84515	149.64	0.0000
Residual	3.88541	120	0.0323784		
Total (Corr.)	32.9563	126			

R-squared = 88.2104 percent  
R-squared (adjusted for d.f.) = 87.621 percent  
Standard Error of Est. = 0.17994  
Mean absolute error = 0.128501  
Durbin-Watson statistic = 2.0155

Use the right mouse button to select options

Inicio PlayCDA AP Practica 5. Vari... TAREA 5 VARI... STATGRAP... 21:28

Puede verse como efectivamente las factorías denominadas *estrella* tienen un coste menor ya que la variable cualitativa es significativa ( $t=-5.8$ ). La ecuación del modelo será:

$$\log(\text{coste de produccion}) = 0.975214 - 0.283898 \cdot \text{estrella} + 0.26584 \cdot \log(\text{salarios}) + 0.137163 \cdot \log(\text{otros costes fijos}) + 0.122714 \cdot \log(\text{energia}) + 0.0936721 \cdot \log(\text{materia prima}) + 0.0975135 \cdot \log(\text{maquinaria})$$

Cuando una factoría pertenece al grupo de las denominadas *estrella* la variable dicotómica generada toma el valor 1 y 0 en caso contrario, por lo que el nuevo término introducido en el modelo supone una disminución de la variable respuesta para las factorías *estrella* ( $-0.28 \cdot 1$ ).

*En las factorías estrella, el coste de producción será 0.28 uds menor que en otra factoría no estrella con iguales valores del resto de las variables.*

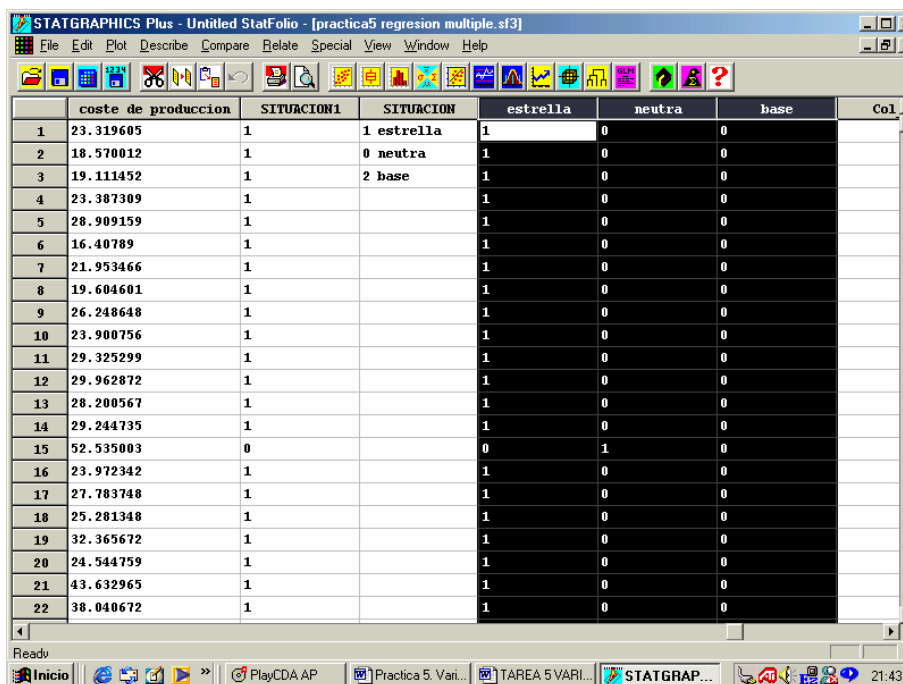
### Ejercicio:

Analizar si el coste de producción es diferente en las factorías base (comparado con el resto de sucursales).

#### 4. Regresión con variables politómicas

Si introducimos en el modelo una única variable dummy en el modelo, estamos comparando las observaciones del grupo analizado con el resto de observaciones. Cuando las observaciones pueden dividirse en varios grupos, puede ser de interés compararlos entre sí. Para ello, se introducen en el modelo variables politómicas.

En el ejemplo analizado, se han generado tres variables dummies, una por cada grupo analizado.



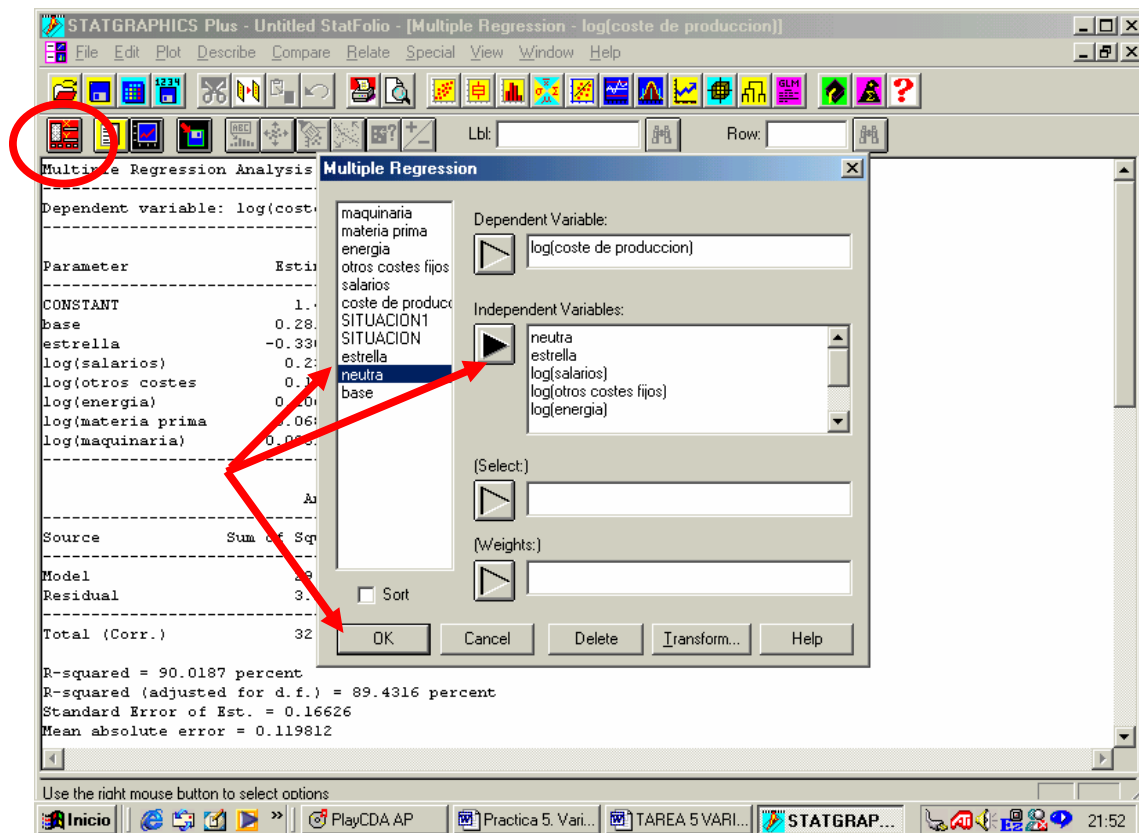
	coste de produccion	SITUACION1	SITUACION	estrella	neutra	base	Col
1	23.319605	1	1 estrella	1	0	0	
2	18.570012	1	0 neutra	1	0	0	
3	19.111452	1	2 base	1	0	0	
4	23.387309	1		1	0	0	
5	28.909159	1		1	0	0	
6	16.40789	1		1	0	0	
7	21.953466	1		1	0	0	
8	19.604601	1		1	0	0	
9	26.248648	1		1	0	0	
10	23.900756	1		1	0	0	
11	29.325299	1		1	0	0	
12	29.962872	1		1	0	0	
13	28.200567	1		1	0	0	
14	29.244735	1		1	0	0	
15	52.535003	0		0	1	0	
16	23.972342	1		1	0	0	
17	27.783748	1		1	0	0	
18	25.281348	1		1	0	0	
19	32.365672	1		1	0	0	
20	24.544759	1		1	0	0	
21	43.632965	1		1	0	0	
22	38.040672	1		1	0	0	

Para analizar la pertenencia a los tres grupos, no es posible introducir las tres variables en el modelo, ya que en este caso el modelo colapsaría. Esto se debe a que la matriz de diseño  $X'X$  no sería invertible por ser la suma de las tres variables cualitativas una columna de unos, y por tanto linealmente dependiente con la columna de la constante.

Para evitar esta colinealidad perfecta tomaremos uno de ellos como referencia (su variable cualitativa no se incorpora al modelo) y para el resto (cuyas variables cualitativas si están en el modelo) se estimará el efecto adicional que tiene la pertenencia al grupo sobre la variable respuesta.

En los datos del ejemplo, tomaremos las factorías *base* como grupo de referencia y estimaremos el diferente coste de las factorías *estrella* y *neutra* respecto al grupo *base*. Para ello, introducimos las variables *estrella* y *neutra* en el modelo.





Obtendremos:

STATGRAPHICS Plus - Untitled StatFolio - [Multiple Regression - log(coste de produccion)]

File Edit Plot Describe Compare Relate Special View Window Help

Multiple Regression Analysis

Dependent variable: log(coste de produccion)

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	1.76685	0.279109	6.33002	0.0000
neutra	-0.281246	0.0605716	-4.6432	0.0000
estrella	-0.612057	0.0837128	-7.31139	0.0000
log(salarios)	0.23143	0.048282	4.79334	0.0000
log(otros costes fijos)	0.10807	0.0329539	3.27942	0.0014
log(energia)	0.106066	0.0220225	4.81625	0.0000
log(materia prima)	0.068208	0.0143157	4.76456	0.0000
log(maquinaria)	0.0951283	0.0261503	3.63775	0.0004

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	29.6669	7	4.23812	153.32	0.0000
Residual	3.28945	119	0.0276425		
Total (Corr.)	32.9563	126			

R-squared = 90.0187 percent  
R-squared (adjusted for d.f.) = 89.4316 percent  
Standard Error of Est. = 0.16626  
Mean absolute error = 0.119812

Use the right mouse button to select options

Inicio PlayCDA.AP Practica 5. Vari... TAREA 5 VARI... STATGRAP... 21:52

Como las dos variables dummies introducidas son significativas (sus estadísticos t cumplen  $t > 2$ ), también lo serán las diferencias entre grupos, de manera que la ecuación del modelo puede leerse:

$$\begin{aligned} \log(\text{coste de produccion}) = & 1.76685 + 0.281246 \cdot \text{neutra} \\ & - 0.612057 \cdot \text{estrella} + 0.23143 \cdot \log(\text{salarios}) + 0.10807 \cdot \log(\text{otros costes fijos}) \\ & + 0.106066 \cdot \log(\text{energia}) + 0.068208 \cdot \log(\text{materia prima}) + \\ & 0.0951283 \cdot \log(\text{maquinaria}) \end{aligned}$$

El efecto de las variables cualitativas puede explicarse del siguiente modo:

Cuando el resto de las variables permanecen constantes, el logaritmo del coste de producción es 0.28 uds menor en las factorías *neutras* si se las compara con las factorías *base* (recordad que la variable cualitativa *base* no se ha introducido en el modelo, lo que significa que todas las comparaciones deben hacerse respecto a este grupo)

Cuando el resto de las variables permanecen constantes, el logaritmo del coste de producción es 0.61 uds menor en las factorías *estrella* si se las compara con las factorías *base*.

Una forma sencilla de analizar las diferencias es escribir las ecuaciones para cada uno de los grupos:

Grupo Base:

$$\text{Log(coste)}=1.7+0.23\log(\text{salarios})+0.1 \log(\text{otroscoste})+0.1 \log(\text{energia})+0.07 \log(\text{mat. Prima})+0.1 \log(\text{maquinaria})$$

Grupo estrella:

$$\text{Log(coste)}=1.7+0.23\log(\text{salarios})+0.1 \log(\text{otroscoste})+0.1 \log(\text{energia})+0.07 \log(\text{mat. Prima})+0.1 \log(\text{maquinaria}) \textbf{-0.6}$$

Grupo neutro:

$$\text{Log(coste)}=1.7+0.23\log(\text{salarios})+0.1 \log(\text{otroscoste})+0.1 \log(\text{energia})+0.07 \log(\text{mat. Prima})+0.1 \log(\text{maquinaria}) \textbf{-0.3}$$

## 5. Interacciones

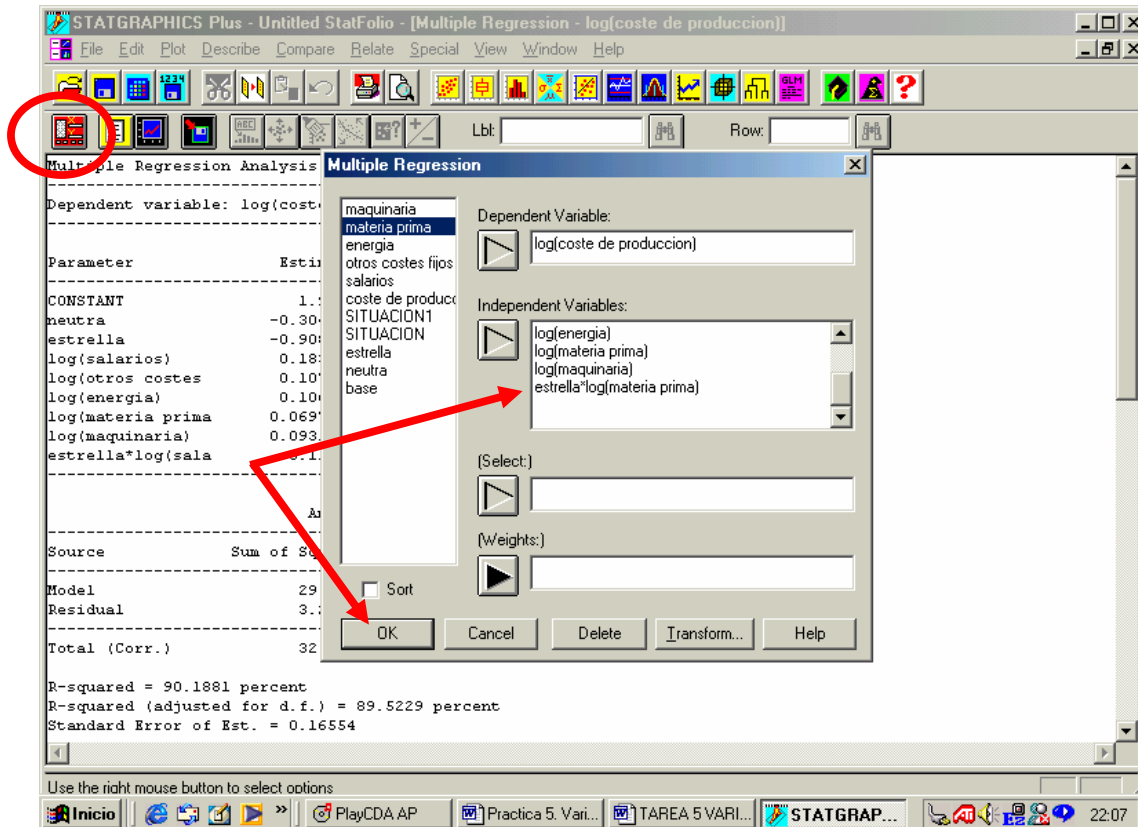
En ocasiones el efecto de una variable explicativa X sobre la variable respuesta puede variar dependiendo de que las observaciones pertenezcan a un grupo u otro. Se dice entonces que se produce interacción entre el grupo y la variable explicativa.

La forma de modelizar dicha interacción es introducir en el modelo una nueva variable que se construye mediante el producto de la variable explicativa y la variable dicotómica correspondiente.

Por ejemplo, para los datos analizados, se sabe que la multinacional analizada ha localizado sus factorías *estrella* en zonas próximas a yacimientos de las materias primas utilizadas. Podemos analizar si el impacto de la materia prima en el coste de producción es diferente dependiendo del emplazamiento de la factoría (es decir, de su pertenencia a uno u otro grupo de los definidos).

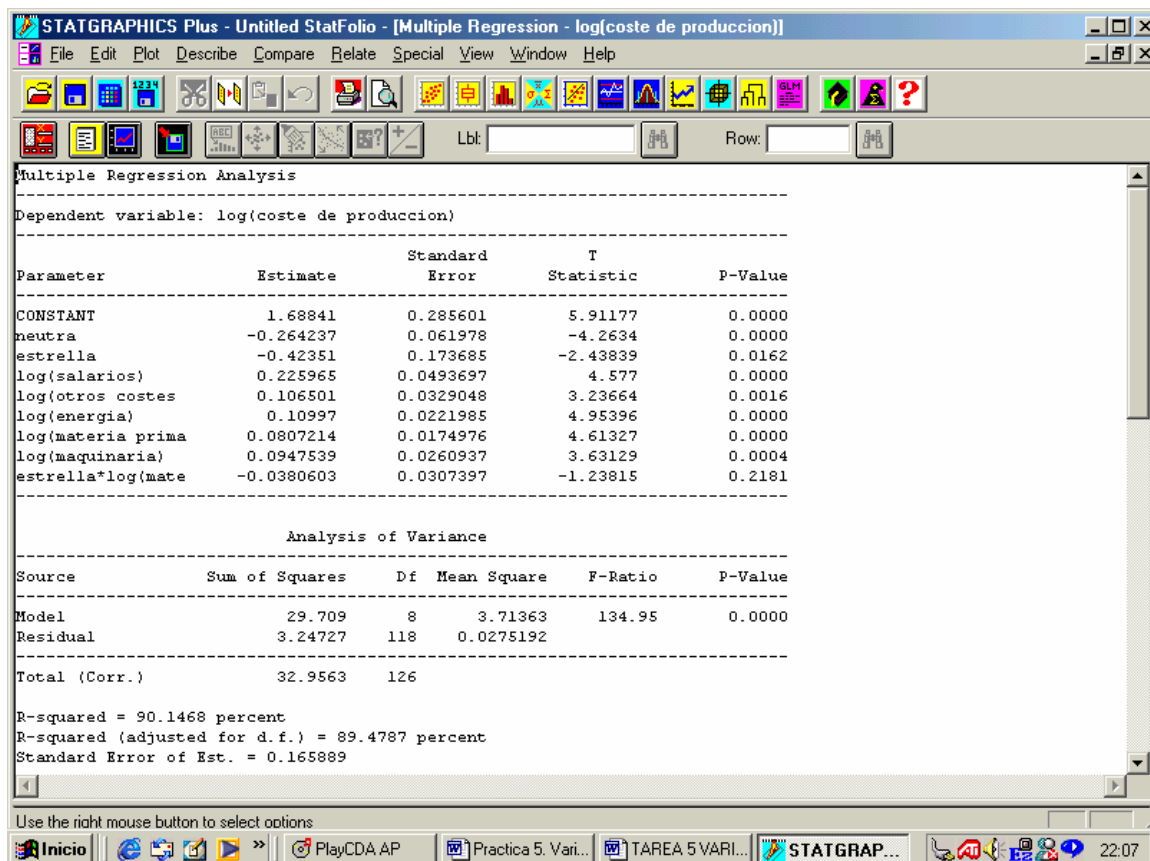
Para ello introduciremos en el modelo una nueva variable que será el producto  
 $\text{Estrella} * \log(\text{materia prima})$

Para ello se hará:



(el producto de las dos variables puede hacerse directamente desde teclado)

Obtendremos:



Donde puede verse que la interacción entre las variables no es significativa ( $t < 2$ ) por lo que la influencia del coste de la materia prima en el coste de producción no es diferente dependiendo del emplazamiento de la factoría. Si hubiera sido significativo, para el análisis de los resultados habría que escribir las ecuaciones de regresión para cada uno de los tres grupos. Hacer el análisis directamente es demasiado complicado.

### Ejercicio:

Analizar si la influencia de los costes energéticos en el coste global de producción depende del emplazamiento de la sucursal. Escribir las ecuaciones de regresión para los tres emplazamientos.