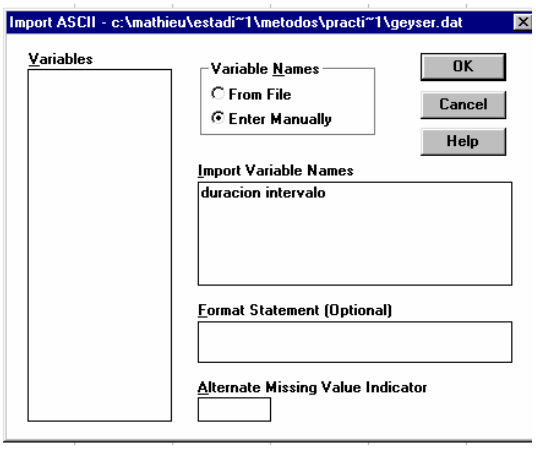
Práctica 3. Explorando datos con Statistix

En esta práctica aprenderemos a explorar un conjunto de datos utilizando el menú Statistics de nuestro programa.

Un geyser es un nacimiento de agua hirviente que de vez en cuando se vuelve inestable y expulsa agua y vapor. El geyser "Old Faithful" en el parque de Yellowstone en Wyoming es probablemente el más famoso del mundo. Los visitantes del parque se acercan al emplazamiento del geyser intentando no tener que esperar demasiado para verlo estallar. Los servicios del Parque colocan un cartel donde se anuncia la próxima erupción. Es por lo tanto de interés estudiar los intervalos de tiempo entre dos erupciones conjuntamente con la duración de cada erupción.
En esta práctica analizaremos los intervalos entre erupciones sucesivas así como la duración de las mismas durante los meses de agosto 1978 y agosto 1979. Se observaron 222 erupciones y los datos de los que disponemos se presentan por pares: (duración de la erupción, intervalo hasta la siguiente). Las unidades de medición son mn.

Para importar los datos en nuestra hoja de cálculo, seleccionamos la opción ***import*** del menu ***File***. A continuación debemos recorrer las carpetas para encontrar el fichero **geyser.dat** (vuestro profesor os dirá en qué carpeta se encuentra). Una vez que hemos localizado el fichero, pulsamos ***OK*** y aparece la ventana siguiente

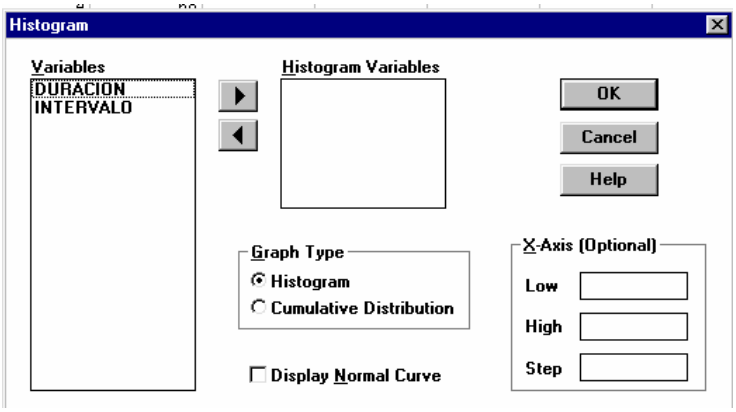


En el cuadro **Variable Names**, escogemos la opción "**Enter manually**" puesto que el fichero de datos no contiene los nombres de la variable, mientras que en el cuadro "**Import Variable Names**" introducimos **Duracion** e **Intervalo** que serán los nombres de nuestras variables. Después de pulsar **OK**, tenemos en nuestra hoja de cálculo los 444 datos.

Podemos empezar con la exploración de los datos: tal como se vio en clase, el primer paso cuando uno dispone de varias variables es empezar por estudiar cada una por separado.

1. Estudio individual de cada variable

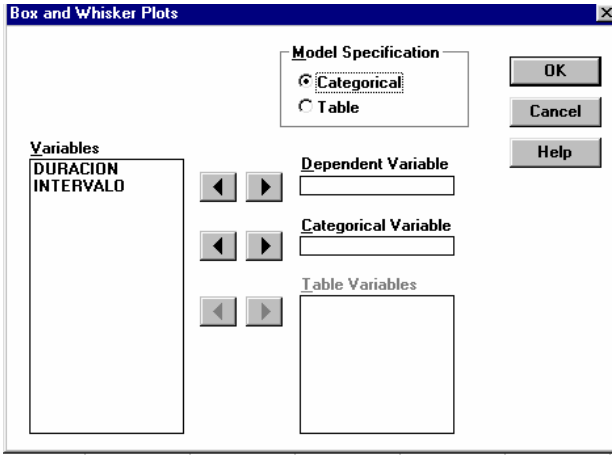
En particular, para hacerse una idea de la distribución de los datos, vamos a realizar un histograma tanto para la duración como para el intervalo. Para ello, seleccionamos la opción **`Statistics->Summary Statistics->Histogram`** y aparece la ventana siguiente:



Por ejemplo, pasamos la variable "***duracion***" desde la izquierda hasta el cuadro "**Histogram variables**". Si no especificamos nada más en el cuadro **X-axis**, el programa realizará de manera automática la elección de clases. Pulsamos por lo tanto **OK**, y aparece el histograma de la duración. Minimizamos la ventana y repetimos los pasos con la variable "**intervalo**".
Describir las características globales de cada histograma:

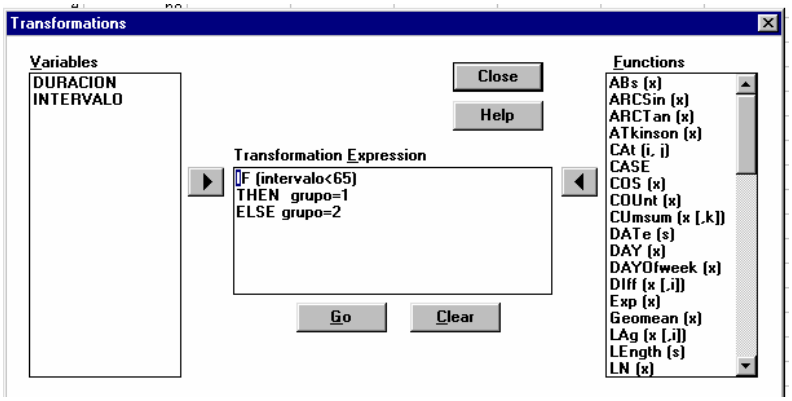
¿Puedes indicar medidas convenientes de centralización y de dispersión de los datos?

Pasamos a realizar un diagrama de caja-bigotes para, por ejemplo, los datos de la duración: Escogemos ***Statistics->Summary Statistics->Box and Whisker plots*** y aparece la ventana siguiente:

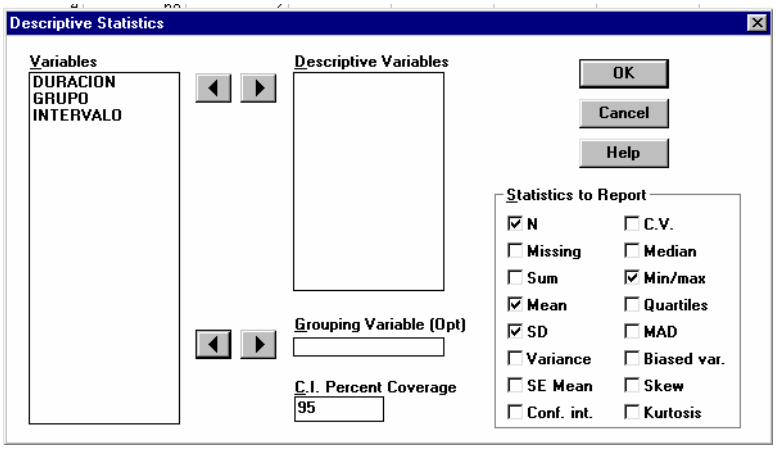


Pasamos la variable ***Duracion*** de la izquierda al cuadro "**Dependent Variable**" y pulsamos **OK**. Como podemos ver, en este caso, el diagrama de cajas y bigotes es un resumen muy expeditivo que oculta la estructura de los datos .

Decidimos estudiar con más detalle los intervalos entre dos erupciones y decidimos separar los dos subgrupos que hemos detectado en el histograma. Empezamos por decidir de un punto de corte, por ejemplo 65mn. Vamos a crear una variable llamada grupo, que valga 1 si el intervalo es menor de 65mn y 2 si es mayor o igual. Para ello, utilizamos la opción **Data->Transformations**, y rellenamos el cuadro "**Transformation expression**" con la siguiente expresión lógica:



Podemos ahora calcular medidas numéricas descriptivas para cada uno de los subgrupos. Para ello, seleccionamos **Statistics->Summary Statistics->Descriptive Statistics**, y nos aparece el cuadro:

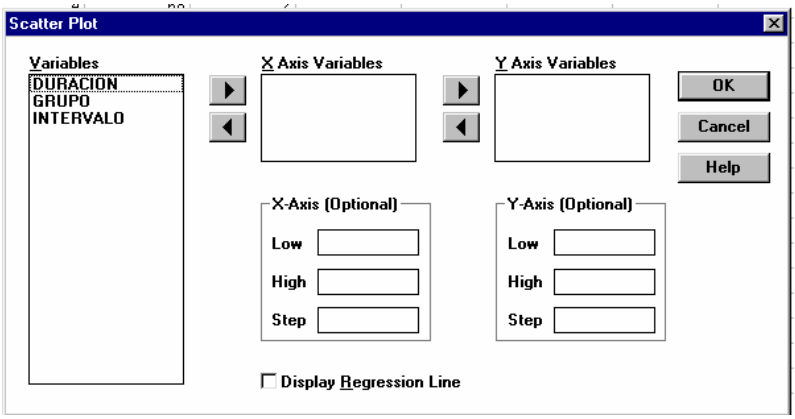


Queremos obtener medidas para dos grupos de valores de "**INTERVALO**". Pasamos por lo tanto la variable INTERVALO al cuadro "**Descriptive variables**" y la variable "**GRUPO**" al cuadro "**Grouping variable (Opt)**". (Si no especificáramos nada en ***Grouping variable***, obtendríamos las medidas para todos los valores de INTERVALO) Seleccionamos en el cuadro "***Statistics to report***", las medidas que nos interesan. Pulsamos **OK** y obtenemos estas últimas.
Para el grupo 1 de INTERVALO:
Media= Desviación típica=
Para el grupo 2 de INTERVALO:
Media= Desviación típica=

Podríamos por supuesto realizar algo parecido para la variable ***DURACION***, decidiendo un punto de corte etc...

2. Estudio conjunto de las dos variables

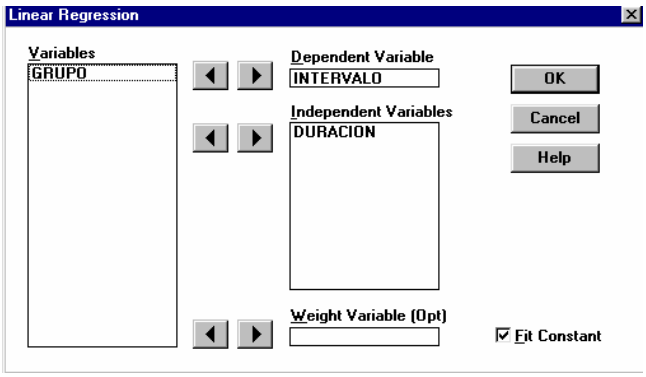
Queremos ahora estudiar la relación que pueda existir entre **DURACION** e **INTERVALO**, en particular, ¿podemos decir que una duración corta implica un intervalo corto o algún tipo de efecto contrario?
Una vez más, lo primero que hay que hacer es conseguir una impresión visual de los datos realizando una nube de puntos. Para ello, seleccionamos la opción ***Statistics-> Summary Statistics -> Scatter plot.***



y pasamos la ***duracion*** al cuadro ***X Axis Variable*** e ***INTERVALO*** al cuadro ***Y Axis Variables***. Pulsamos **OK**, y obtenemos la nube de puntos.

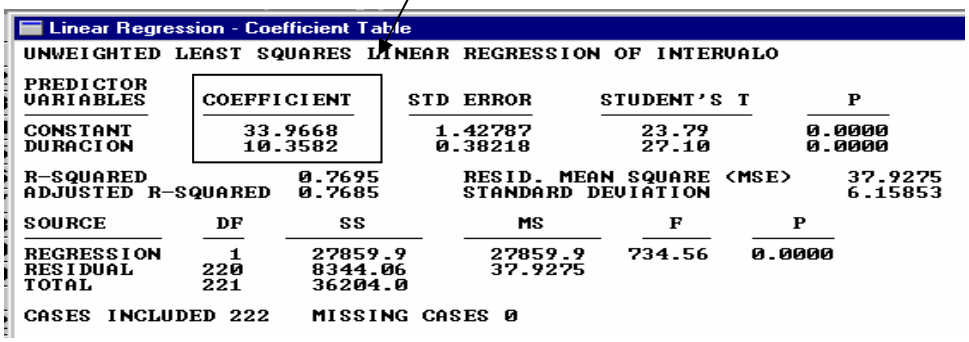
¿Qué tipo de relación existe entre el intervalo de tiempo hasta la siguiente erupción y la duración de la última? ¿Cuál podría ser un modelo teórico?

Si decidimos ajustar una recta a la nube de puntos, podemos conseguir de manera automática los coeficientes de la recta:
Seleccionamos ***Statistics->Linear Models->Linear regression***,



y pasamos ***INTERVALO*** como variable dependiente, y ***DURACION*** como variable independiente. La casilla "***Fit constant***" corresponde a si queremos que calcule la ordenada al origen o si forzamos la recta por el origen. En este caso, la mantenemos activada. Al pulsar **OK**, obtenemos la tabla siguiente:

Coeficientes de la recta de regresión



Linear Regression - Coefficient Table

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF INTERVALO

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	33.9668	1.42787	23.79	0.0000
DURACION	10.3582	0.38218	27.10	0.0000

R-SQUARED	0.7695	RESID. MEAN SQUARE (MSE)	37.9275
ADJUSTED R-SQUARED	0.7685	STANDARD DEVIATION	6.15853

SOURCE	DF	SS	MS	F	P
REGRESSION	1	27859.9	27859.9	734.56	0.0000
RESIDUAL	220	8344.06	37.9275		
TOTAL	221	36204.0			

CASES INCLUDED 222 MISSING CASES 0

Deducimos que podemos proponer como modelo teórico para explicar el intervalo de tiempo entre dos erupciones en función de la duración de la última erupción :

INTERVALO= 33.97+10.36*DURACIÓN.

En cuanto a la bondad del ajuste, se lee el coeficiente de determinación R^2 en frente de "R-SQUARED". En nuestro caso, encontramos un coeficiente de determinación igual a 0.77, lo que indica que nuestro ajuste es aceptable.

Gracias a un modelo como éste, los servicios del parque son capaces de predecir con una precisión satisfactoria, después de una erupción, cuándo será la siguiente.

Práctica 4 : Ajuste por Mínimos Cuadrados.

¡Error! Marcador no definido.

Veamos como resolver el siguiente problema utilizando el programa Statistix:

Se ha realizado un estudio para investigar el efecto de un determinado proceso térmico en la dureza de una determinada pieza. Once piezas se seleccionaron para el estudio. Antes del tratamiento se realizaron pruebas de dureza para determinar la dureza de cada pieza. Después, las piezas fueron sometidas a un proceso térmico de templado con el fin de mejorar su dureza. Al final del proceso, se realizaron nuevamente pruebas de dureza y se obtuvo una segunda lectura. Se recogieron los siguientes datos (Kg. de presión):

¡Error!	1	2	3	4	5	6	7	8	9	10	11
Dureza previa	**182**	**232**	**191**	**200**	**148**	**249**	**276**	**213**	**241**	**480**	**262**
Dureza posterior	**198**	**210**	**194**	**220**	**138**	**220**	**219**	**161**	**210**	**313**	**226**

a) Calcular la dureza media antes y después del proceso. Así como las desviaciones típicas en cada caso.
b) Realizar un diagrama de caja bigotes para la dureza previa y la dureza posterior.
c) Estudiar el ajuste de mínimos cuadrados del nivel posterior con respecto al nivel previo de dureza.
d) Estudiar la precisión del ajuste anterior.

Resolución:

Empezamos por definir gracias al menu ***Data->insert->Variables***, dos nuevas variables **Previa** y **posterior**. A continuación introducimos los valores de estas variables.

a) Para calcular medidas numéricas asociadas a cada variable, utilizamos las nociones vistas en la práctica anterior: seleccionamos la opción

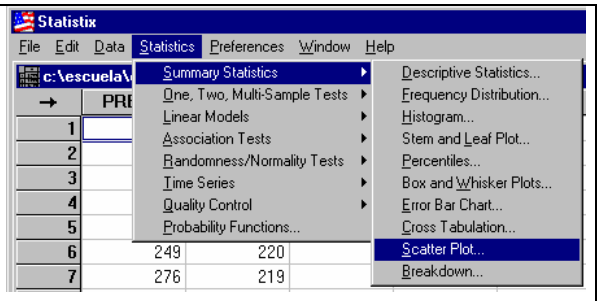
Statistics->Summary Statistics->Descriptive Statistics.

b) Igualmente, para realizar un diagrama de cajas-bigotes, seleccionamos, tal como lo vimos en la última práctica, la opción ***Statistics->Summary Statistics->Box and Whisker Plots***. Especificamos el modelo en forma de tabla ("***Table***") y pasamos las variables X e Y al cuadro "***Table variables***".

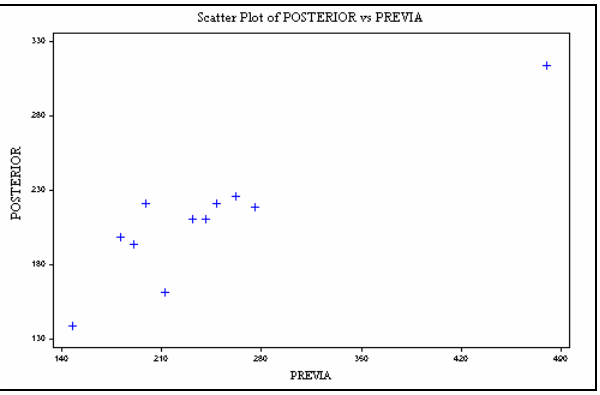
c)

Representaremos en primer lugar el gráfico de dispersión o también llamado nube de puntos con el fin determinar si existe una cierta tendencia lineal. Para ello seleccionaremos:

Statistics->Sum. Statistics->Scatter Plot.

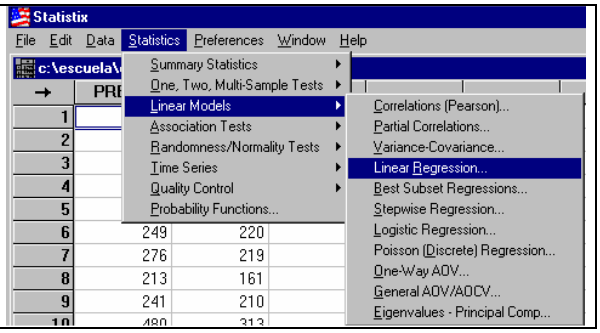


Como se observa en la gráfica, existe un punto demasiado alejado (se corresponde con los resultados de la pieza 10, (480,313)) que en principio puede dar un resultado engañoso sobre la dependencia lineal de ambas características.



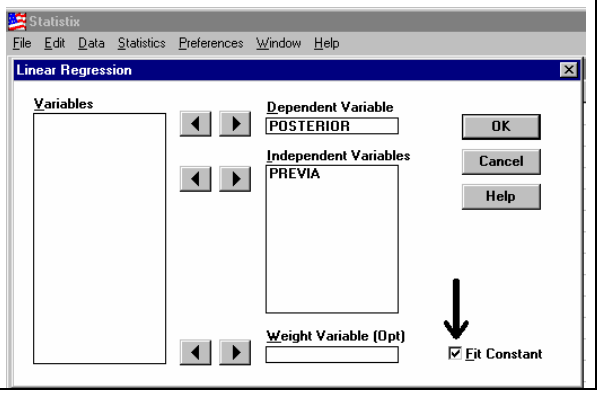
En cualquier caso, pasemos a calcular la ecuación de la recta de regresión, para lo cual seleccionaremos:

Statistics-> Linear models-> ->Linear Regresion



Indicaremos como variable independiente la variable **Previa** y como dependiente la variable Posterior puesto que pretendemos estudiar la dureza **Posterior** en función de la dureza previa.

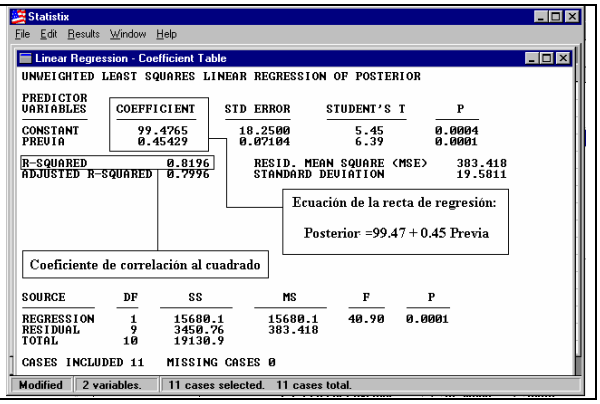
Observar que en la parte inferior aparece una casilla con el indicador Fit Constant**. Esta casilla** debe estar marcada **cuando se ajusta una recta de la forma y=ax+b, pero se debe desactivar para una recta forzada por el origen: y=ax.**



Como podemos observar, Statistix nos proporciona la ecuación de la recta ajustada y el coeficiente de correlación al cuadrado:

$$Y = 99.47 + 0.45\ X$$
$$R^2 = 0.81$$

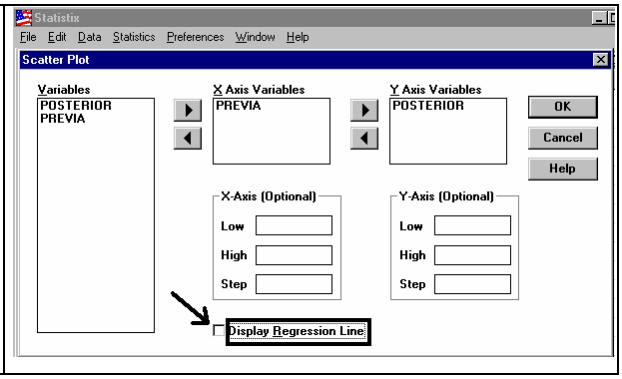
junto con otra información que comentaremos en posteriores prácticas.



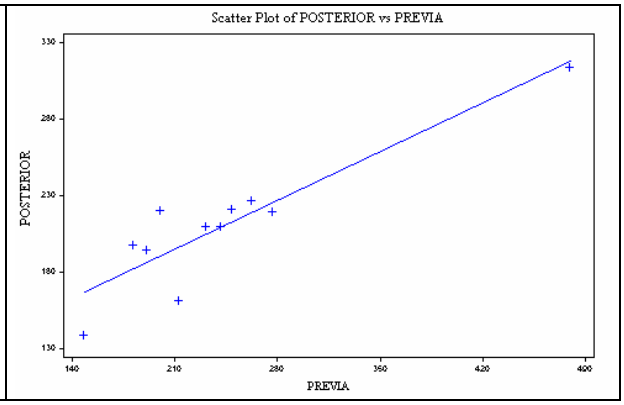
Si recordamos, al seleccionar:

Statistics-> Summary Statistics->Scatter Plot.

tenemos la posibilidad de presentar junto con el diagrama de dispersión la recta de regresión con el fin de observar si existe algún valor que presente un gran residuo.



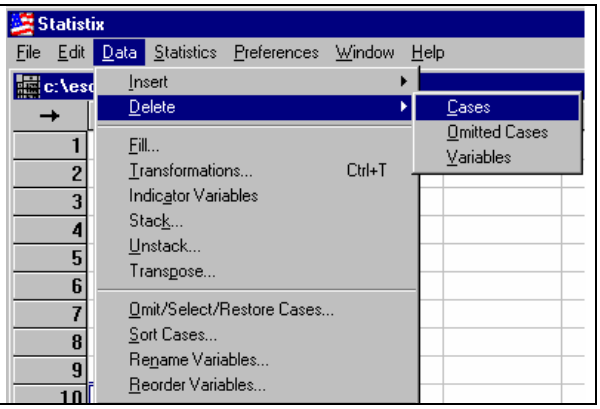
La gráfica que obtenemos es:



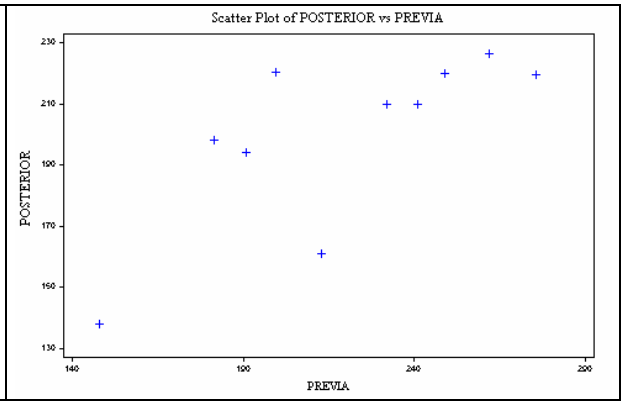
Veamos que ocurre si eliminamos el dato número 10, el nuevo gráfico de dispersión no presenta una clara tendencia lineal.

Para eliminar el dato seleccionaremos:

<u>Data -> Delete -> Cases</u>

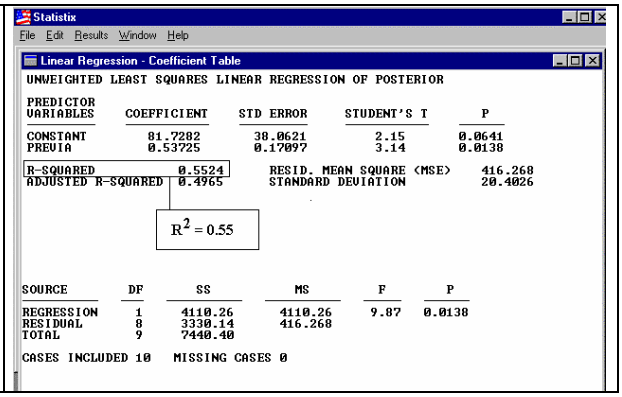


Una vez eliminado dicho valor, el gráfico de dispersión indica que no parece existir una clara tendencia lineal entre los puntos.



Como se observa, si realizamos el ajuste obtenemos un coeficiente de correlación bastante bajo:

R^2=0.55



Conclusión:

Si el valor correspondiente a la pieza 10 no se descarta, se obtiene un coeficiente de determinación de 81%, lo que indica un buen ajuste lineal. Sin embargo, si el valor obtenido para la pieza 10 resulta erróneo y lo descartamos, ,se obtiene un coeficiente de determinación de 55%, lo que pone en duda la validez de un ajuste lineal entre los resultados obtenidos antes y después del proceso de templado.

Ejercicios propuestos
Resolución con Statistix

1) Las materias primas empleadas en la producción de una fibra sintética son almacenadas en un lugar en donde no se tiene control de la humedad. La siguiente tabla refleja en porcentajes la humedad relativa del almacén X y la humedad observada en la materias primas Y durante un estudio que tuvo lugar durante 12 días.

X	41	53	59	65	71	78	50	65	74
Y	1.6	13.6	19.6	25.6	31.6	33.2	14.7	21.2	28.3

a) Realizar un ajuste de mínimos cuadrados entre ambas variables.
b) Estudiar la precisión del ajuste en función del valor obtenido por el coeficiente de correlación, representar gráficamente la recta hallada y comentar los resultados.

2) Con el fin de determinar si existe relación entre la cantidad de polímeros de látex incluida durante el proceso de mezclado de cemento Portland y su resistencia adhesiva a tensión, una empresa encargada de realizar certificaciones de obras toma una muestra de tamaño 10, obteniendo los siguientes resultados

X	13.5	11.0	13.0	11.2	12.0	13.2	12.0	13.5	11.2	13.0
Y	17.5	16.6	17.2	16.6	17.0	17.3	16.9	17.3	16.8	17.1

a) Calcular la media y varianza asociada a cada una de las variables.
b) Calcular la covarianza existente entre ambas variables así como el coeficiente de correlación.
c) Realizar un ajuste por mínimos cuadrados de la resistencia respeto a la cantidad de polímeros añadida en la mezcla.

3) La hidrólisis de un cierto éster tiene lugar en medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial de $3.^{10\text{-}2}$ M del éster, se han medido las concentraciones del mismo a diferentes tiempos obteniéndose los resultados siguientes.

T (mn)	3	4	10	15	20	30	40	50	60	75	90
C 10^{-3}(M)	25.5	23.4	18.2	14.2	11	6.7	4.1	2.5	1.5	0.7	0.4

a) Realice una nube de puntos de las dos variables. ¿ Le parece adecuado un modelo lineal para escribir este conjunto de datos?
b) Defina una nueva variable Y' que sea Y'=ln (concentración) y realizar la nube de puntos Y' en función de t.
c) Realizar un ajuste por mínimos cuadrados de Y' sobre t con un modelo del tipo: y=ax+b. ¿Cuál es el modelo teórico que propone para C en función del tiempo?
d) Nos dan la información adicional de que se sabe con exactitud que la concentración inicial para T=0 era igual a 30.10^{-3}M. ¿Cómo podemos incluir esta información en nuestro modelo?