Descriptiva Bivariante con STATGRAPHICS CENTURION
-Dependencia lineal y Regresión-

Fichero de datos empleado: VelVientos730.sf3

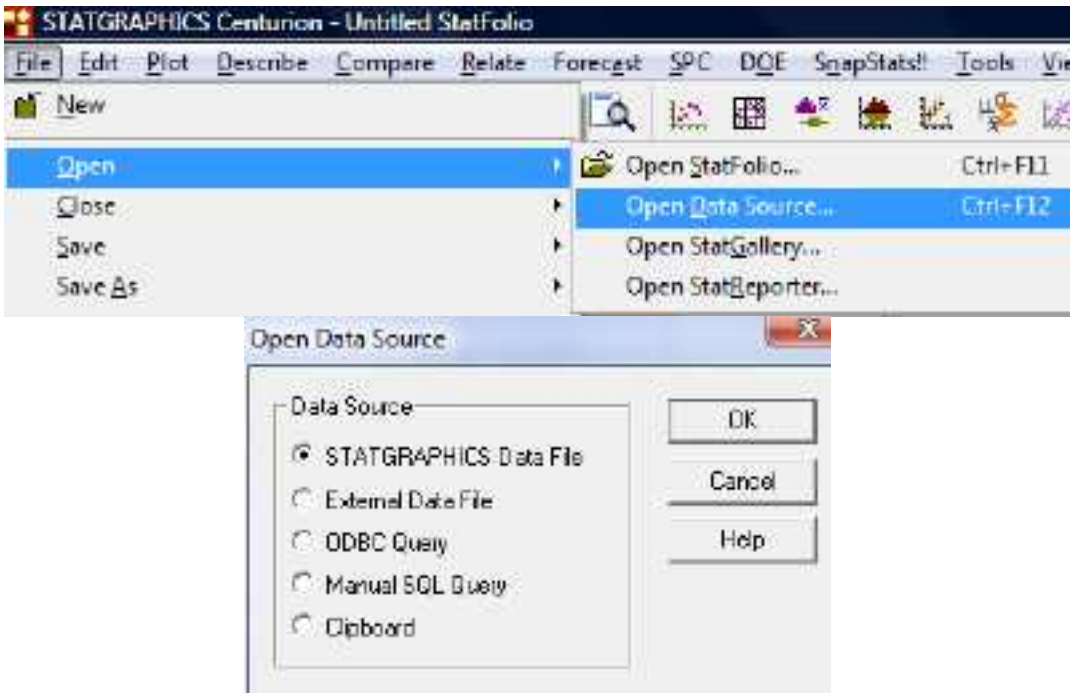
1. Introducción

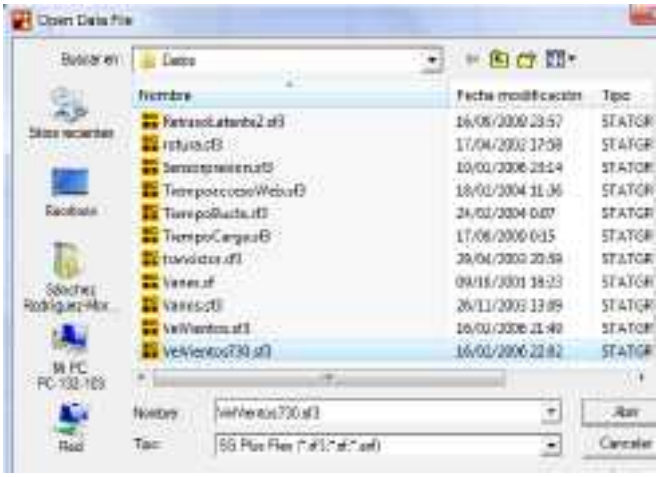
En este documento se analizarán, utilizando Statgraphics Centurion XVI, dos variables observadas simultáneamente. Se analizará su dependencia lineal y se construirá la recta de regresión que ayude a predecir una variable a partir de la segunda.

Los datos que se tienen son los registros de velocidades de viento de dos anemómetros colocados en dos parques eólicos cercanos. El fichero VelVientos.sf3 contiene el registro de 730 horas, donde en cada hora se tiene la velocidad del viento registrada en cada parque. Las velocidades, en metros por segundo (m/s), de cada parque se encuentran en las variables Parque1 y Parque2 respectivamente.

Se quiere disponer de un sistema informático que registre las velocidades del viento en esos parques en tiempo real. Esa información es muy importante para poder gestionar la producción energética del parque y también para detectar errores de funcionamiento de los aerogeneradores. El sistema informático que se instalará es muy costoso, pues requiere una red donde algunas etapas usan transmisión por microondas, calibraciones periódicas, y personal y procedimientos que monitoricen las transmisiones de los ficheros. Por esta razón se decide realizar esa instalación sólo para el Parque1. El objetivo final que se pretende con el análisis de los datos es utilizar las mediciones de viento del Parque1 para predecir las del Parque2 mediante una recta de regresión, y ahorrarnos así duplicar el coste del sistema.

Lo primero que hacemos es leer ese fichero de datos.





VelVientos730.sf3

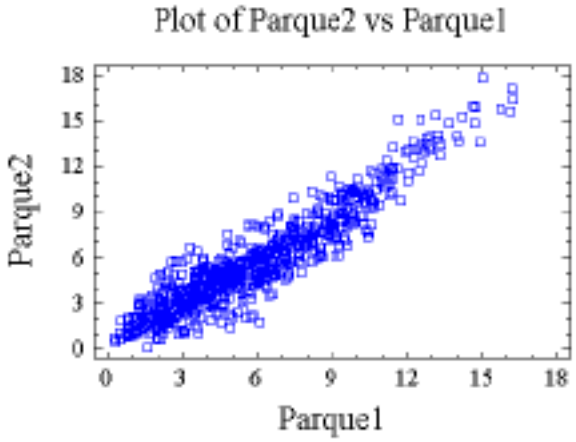
	Parque1	Parque2
	Velocidad viento m/s	Velocidad viento m/s
1	3,50	4,03
2	2,19	2,75
3	1,93	1,99
4	2,08	1,92
5	2,86	2,28

2. Análisis Gráfico

Veamos el gráfico de dispersión de estas dos variables. El Statgraphics permite hacer gráficos de dispersión en varios lugares. El lugar más sencillo es el siguiente: Plot/Scatterplots/X-Y Plot

Como nuestro objetivo es usar al Parque1 como variable explicativa, y al Parque2 como variable respuesta, llamaremos X=Parque1 e Y=Parque2, aunque a efectos de hacer el gráfico esa distinción sea arbitraria

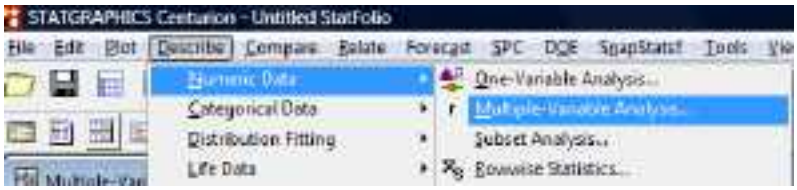
El gráfico que resulta es



donde se aprecia que la relación entre ambas variables es lineal y muy fuerte. Parece entonces sensato utilizar una recta de regresión para predecir Y en función de X. El que las distribuciones de ambas variables sean parecidas sin duda ayuda a que la relación sea mayor.

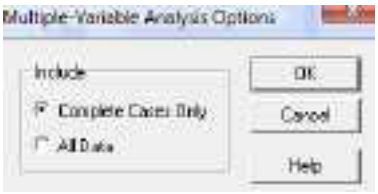
3. Medidas características bivariantes

Para buscar las medidas características que resuman esta relación lineal vamos a



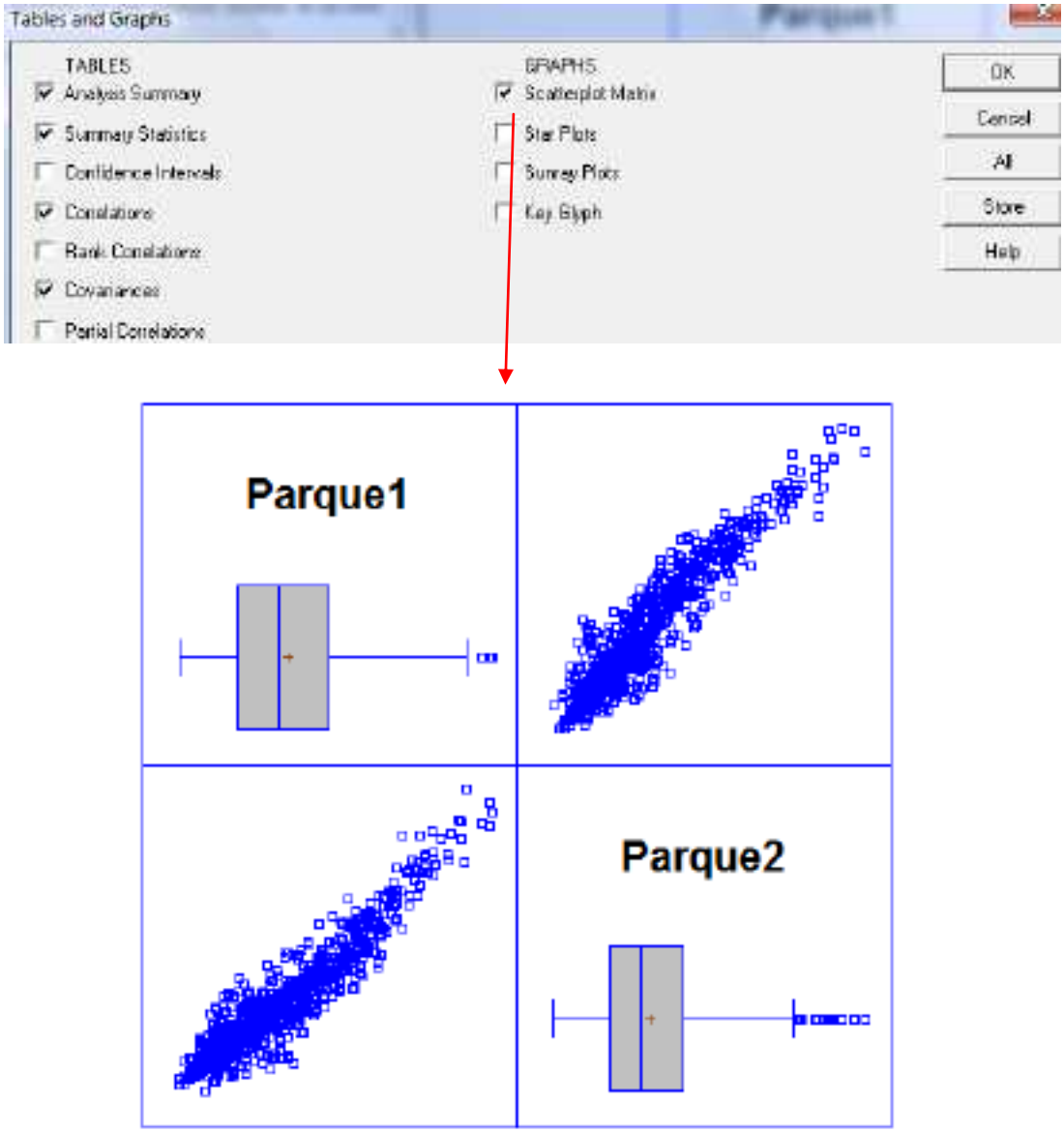
y allí seleccionamos nuestras dos variables

Aparece entonces la siguiente ventana



Si seleccionamos la primera opción, se eliminarán aquellas filas que tengan algún valor ausente en alguna de las variables. Si se selecciona la segunda opción, se usarán todos los datos posibles. Por ejemplo, si tenemos tres variables V1, V2 y V3, y en V1 falta el dato de la fila 12, bajo la opción primera se elimina esa fila para cualquier análisis. Con la segunda opción, la fila 12 se utiliza en aquellos cálculos en los que no intervenga V1; por ejemplo, para calcular la correlación entre V2 y V3.

Las opciones gráficas de esta sección muestran también un diagrama de dispersión junto con un box-plot de cada variable. Las opciones numéricas que nos interesan son la matriz de covarianzas y, sobre todo, la matriz de correlaciones.



La matriz de covarianzas es:

Covariances

	Parque1	Parque2
Parque1	10,5057 (730)	9,84153 (730)
Parque2	9,84153 (730)	10,5948 (730)

De la información de esta matriz podríamos ya deducir las correlaciones e incluso el coeficiente de regresión. Por ejemplo, la correlación entre ambos parques será

$$corr(x,y) = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{9.84153}{\sqrt{10.5057}\sqrt{10.5948}} = 0.9328$$

Esta correlación coincide con el resultado que suministra el Statgraphics

Correlations

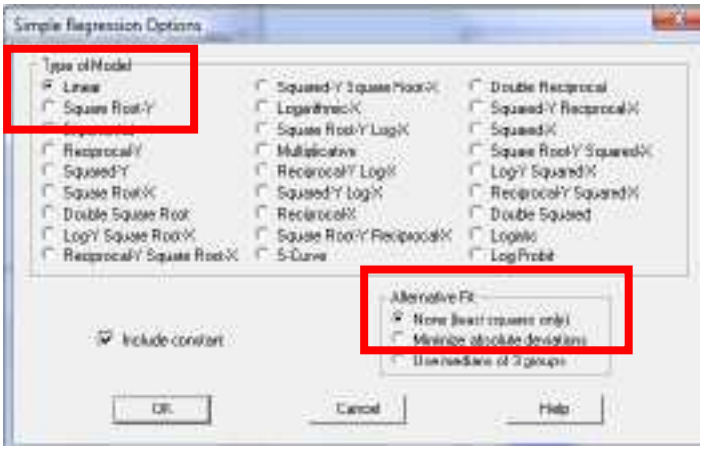
	Parque1	Parque2
Parque1		0,9328 (730) 0,0000
Parque2	0,9328 (730) 0,0000	

Un gráfico de dispersión tan lineal y una correlación tan alta harán que la recta de regresión vaya a ser muy precisa.

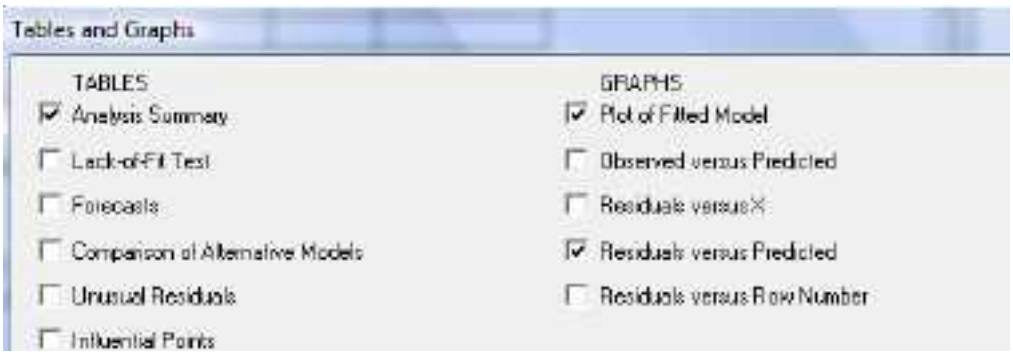
4. La recta de regresión

Para calcular la recta de regresión, también llamada de regresión simple (por tener una sola variable explicativa), nos vamos a Relate/One Factor/Simple Regression

y allí seleccionamos las variables implicadas. Variable respuesta=Y=Parque 2; Variable explicativa=X=Parque1. El Statgraphics Centurion nos muestra una ventana para que elijamos la función que queremos ajustar a los datos. Seleccionamos el modelo lineal. También marcamos que sólo queremos el ajuste por el método de mínimos cuadrados



La técnica de regresión tiene muchos más implicaciones teórico-prácticas que las que se exponen en este documento, por lo que la mayoría de las opciones que nos muestra el Statgraphics no nos son de ayuda. En lo que respecta a opciones numéricas, seleccionamos sólo el resumen de los resultados



La ecuación que queremos estimar es la recta de mínimos cuadrados

$$\hat{y}_i = a + bx_i$$

donde

$$b = \frac{\text{cov}(x, y)}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

El cálculo de estos valores a,b que nos proporciona el programa es:

Coefficients

	Least Squares	Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
Intercept	0,197868	0,0891091	2,22051	0,0264
Slope	0,936776	0,0134105	69,8538	0,0000

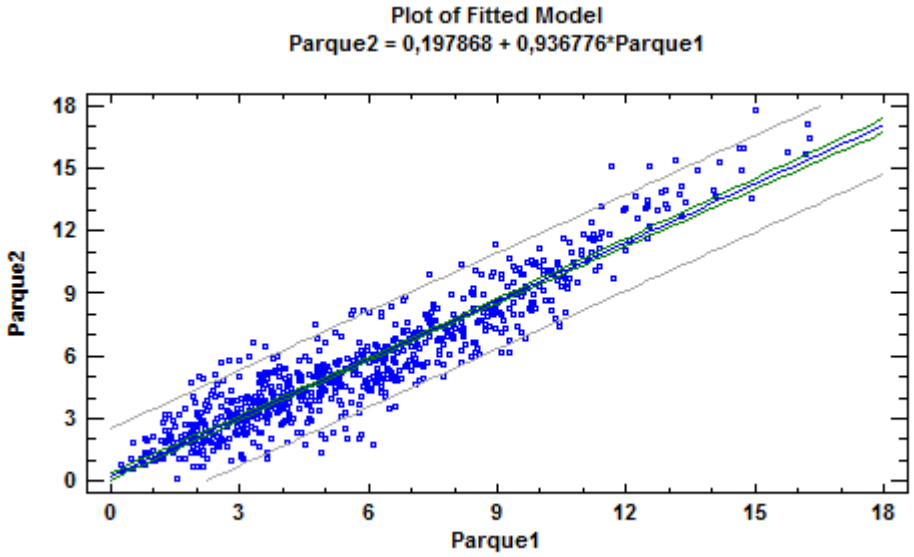
a

b

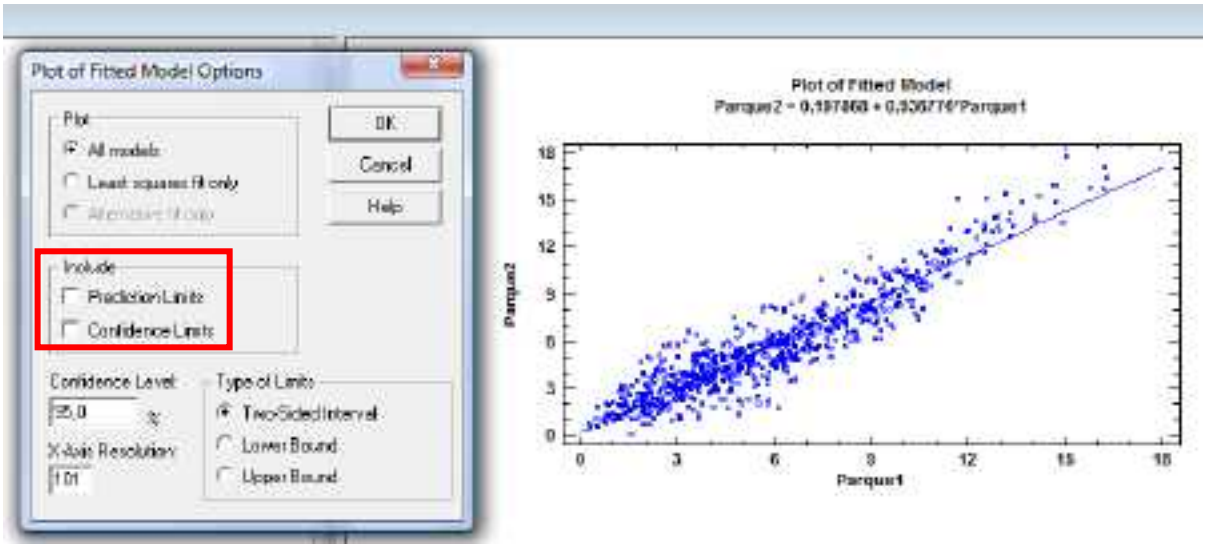
donde sólo nos interesan los valores correspondientes a la columna *Least Squares Estimate*. La pendiente de la recta, es decir, el parámetro **b**, es *Slope*, y el punto de corte cuando x=0 , es decir, el parámetro **a**,es el *Intercep*. Nuestra recta de regresión es entonces

Velocidad Prevista en Parque2=0.198+0.937x(Velocidad del Parque1)

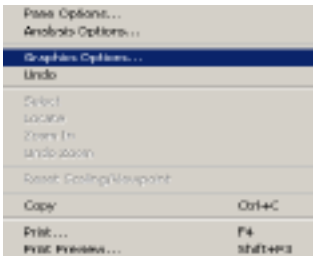
La recta de regresión aparece dibujada en las opciones gráficas (*Plot Fitted Model*)



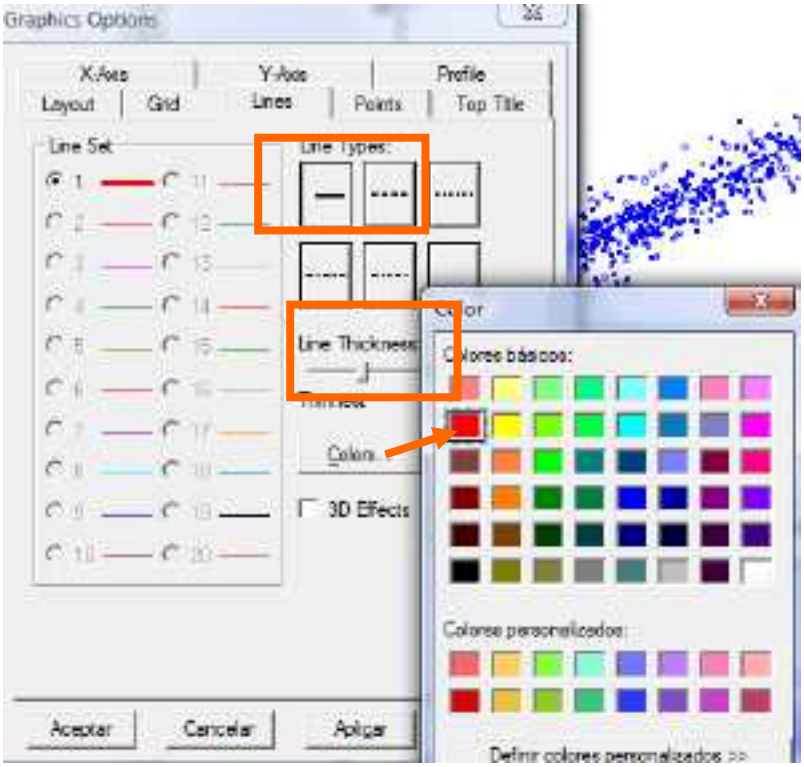
El gráfico que resulta tiene unas líneas auxiliares que al nivel en que estamos manejando la recta de regresión en este documento tampoco nos son de utilidad. Para quitarlas, nos colocamos en el gráfico y pulsamos el botón derecho del ratón. Accedemos así a *Pane Options*. Allí accedemos a las opciones de este gráfico, donde eliminamos las curvas que no nos interesan.



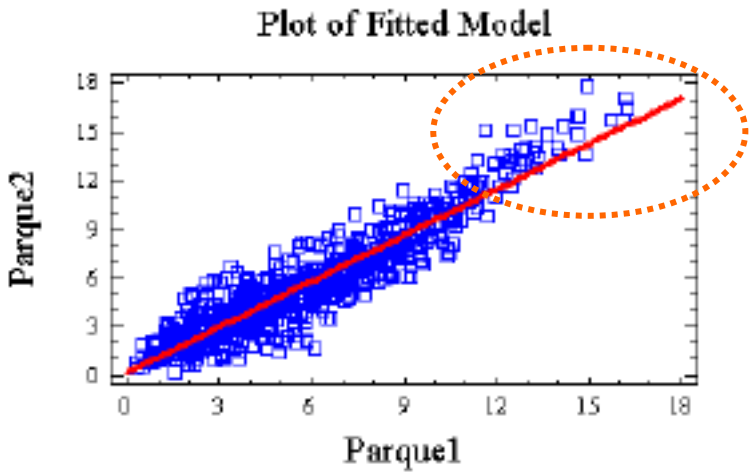
Para mejorar la visualización de la recta de regresión nos colocamos de nuevo sobre el gráfico, pulsamos el botón derecho del ratón y seleccionamos Graphics Options



y allí modificamos el grosor y el color de la línea



El gráfico que resulta es ahora más claro.

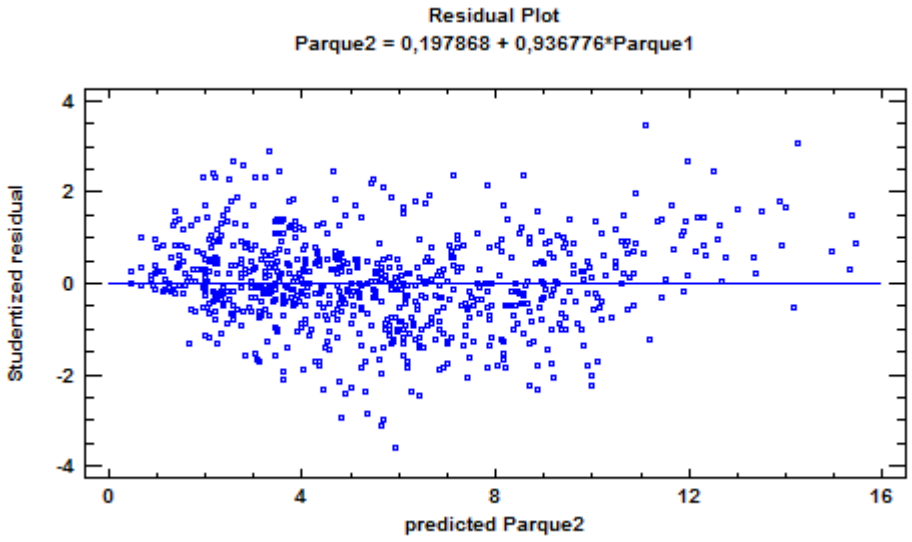


La recta subestima la velocidad del parque2 en esta zona

La recta junto con la nube de puntos muestra que, si bien el ajuste es bastante bueno en casi todo el rango de velocidades, a velocidades altas del Parque1, la recta subestima la velocidad del Parque2. El coeficiente de determinación nos lo proporciona el Stagraphics

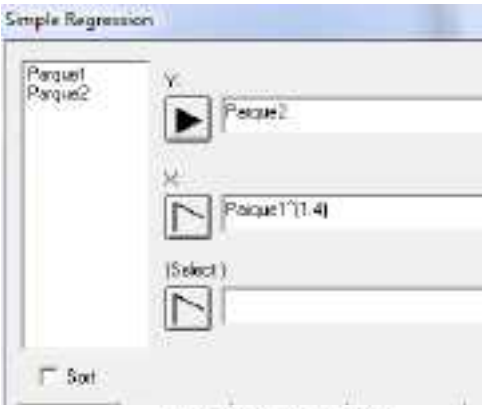
Correlation Coefficient = 0,932832
R-squared = 87,0175 percent

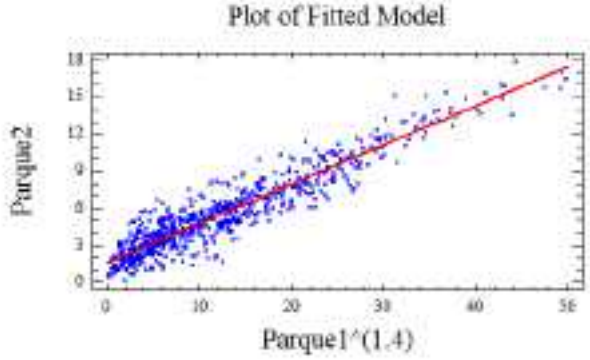
Por lo tanto, si asumimos la relación lineal entre ambas variables como satisfactoria, la velocidad del viento en el parque 2 viene explicada en un 87% por la del parque 1, por lo tanto es bastante buen predictor. La diagnosis de la linealidad del modelo muestra, sin embargo, que el modelo es muy deficiente. Para hacer la diagnosis de la linealidad del modelo realizamos el gráfico de residuos frente a valores previstos, que está entre las opciones gráficas (*Residual versus Predicted*):

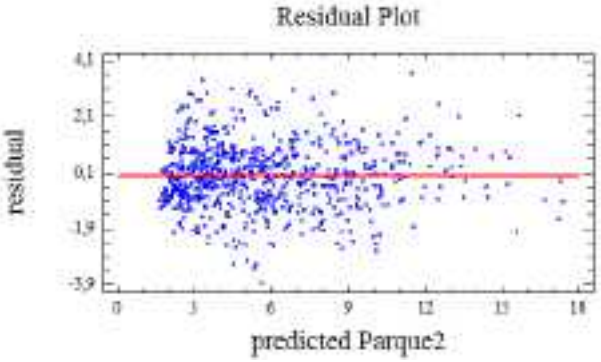


Este gráfico de residuos no es del todo convincente pues, como se veía más arriba, muestra que a valores altos no se ajusta bien. El gráfico de residuos frente a valores previstos muestra una curvatura que señalaría falta de linealidad en los datos. Esta falta de linealidad también puede verse en el gráfico XY anterior con la recta superpuesta, pero es mucho más difícil de apreciar.

La no linealidad que presentan los datos podría corregirse utilizando una transformación del tipo x^c, con c>1. De esta forma expandiríamos los datos del eje X de forma tal que los valores más altos se alejasen más, corrigiendo la curvatura que presentan los datos. La figura siguiente muestra los resultados de la transformación x^(1.4) que parece solucionar el problema.







El modelo final es:

```
Parque2 = 1,58746 + 0,317735*Parque1^(1.4)
```