

ACTIVIDAD 5: Correlación y Regresión Lineal

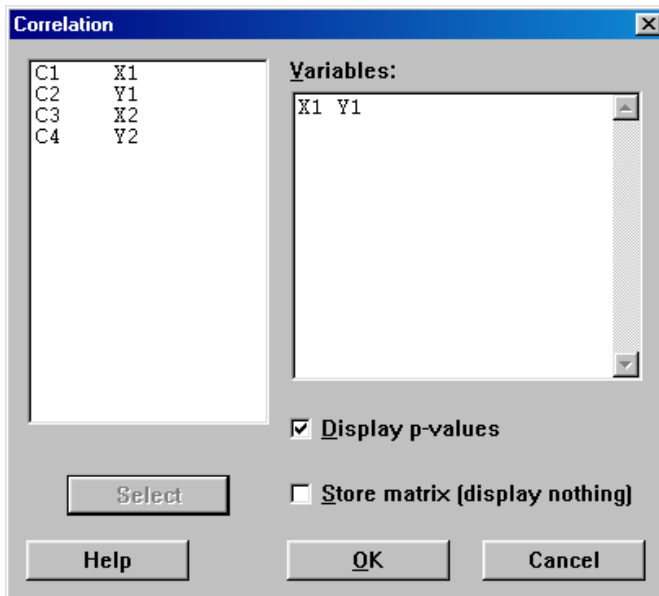
CASO 5-1: RELACIONES ENTRE VARIABLES

A continuación se muestran cuatro variables y seis valores (observaciones) asociados a cada una de ellas:

X1	Y1	X2	Y2
1	4	1	1
2	5	2	4
3	6	3	7
4	7	4	7
5	8	5	4
6	2	6	1

1. Calcular la correlación entre X1 e Y1. ¿Crees que hay algún tipo de relación entre ambas variables? Dibuja la nube de puntos asociada. ¿Qué opinas ahora?

☞ Seleccionamos *Stat > Basic Statistics > Correlation* :

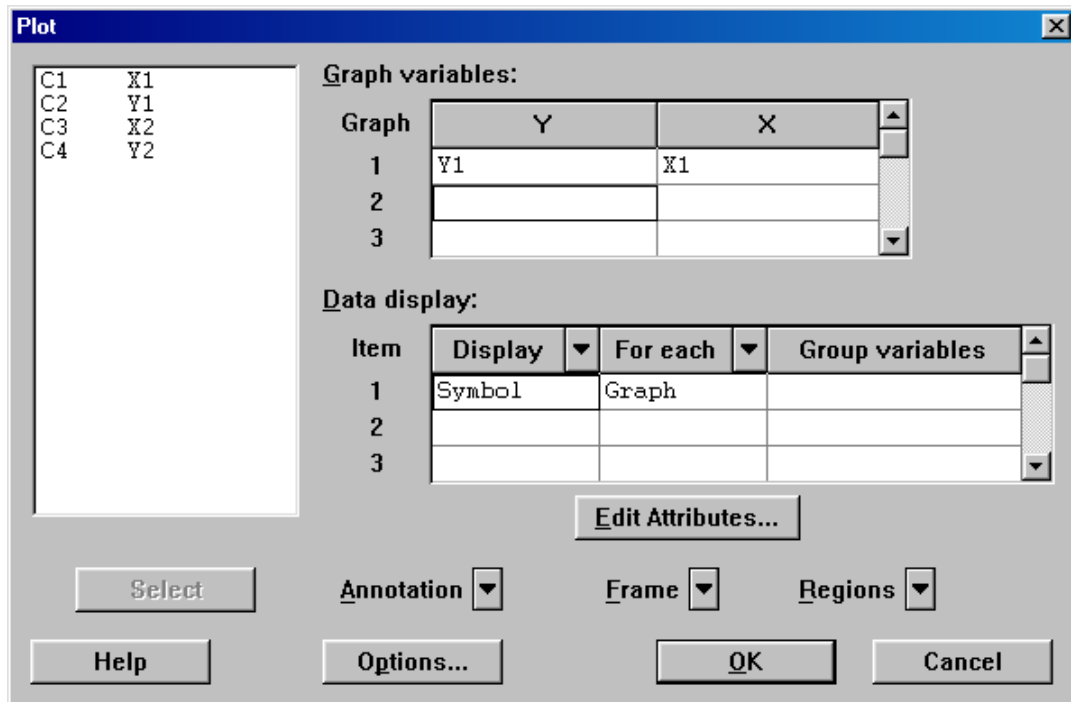


Correlations (Pearson)

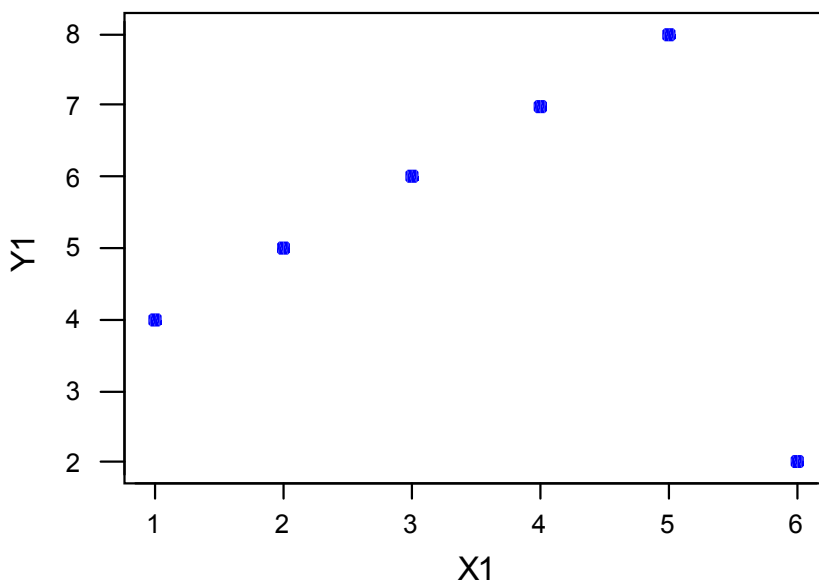
Correlation of X1 and Y1 = 0,000; P-Value = 1,000

En principio, dado que el coeficiente de correlación de Pearson (r) es igual a 0,000 es lógico pensar que no hay relación lineal entre ambas variables.

☞ Seleccionamos *Graph > Plot* :



Nube de puntos X1 vs Y1



¡Sorpresa!: viendo la gráfica parece claro que la causa de que r sea 0 es atribuible a la existencia de un “outlier” (punto muy alejado del resto). Notar que si no fuese por dicho “valor extraño”, podríamos decir que la relación entre ambas variables sería lineal (de hecho hubiésemos obtenido un $r = +1$).

2. Repetir el apartado anterior con las variables X2 e Y2.

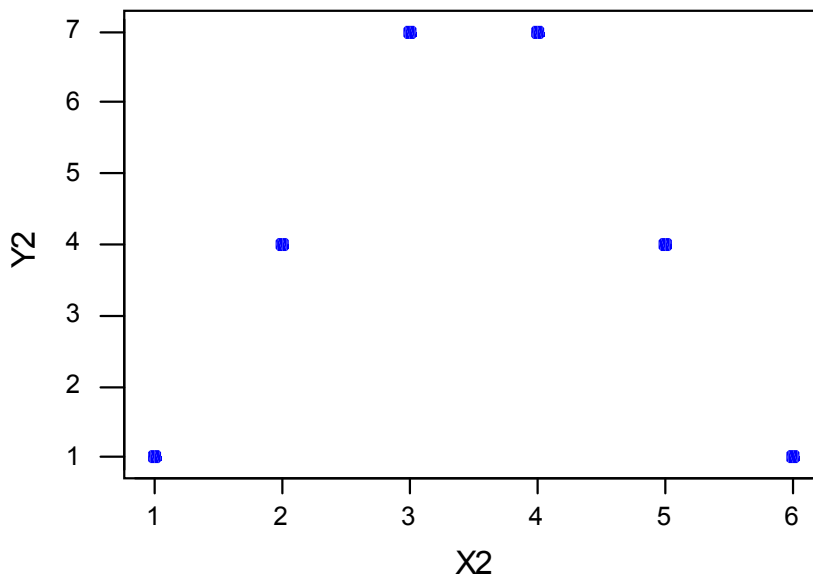
De forma análoga, podemos calcular el nuevo coeficiente de correlación lineal. Obtendremos nuevamente un valor de $r = 0$:

Correlations (Pearson)

Correlation of X2 and Y2 = 0,000; P-Value = 1,000

Si dibujamos la nube de puntos asociada, entenderemos el por qué de tal valor. Como se puede apreciar en el gráfico, existe una relación polinómica no lineal entre ambas variables. Es por ello que $r = 0$ ya que este coeficiente sólo mide la existencia de relaciones lineales.

Nube de puntos de X2 vs Y2



CASO 5-2: EVOLUCIÓN HISTÓRICA DE UN TEST

En un instituto de bachillerato se ha llevado a cabo el siguiente experimento: a lo largo de 15 años (desde 1986 hasta el 2000) se han realizado dos tests a los alumnos del último curso, uno de lengua y otro de matemáticas.

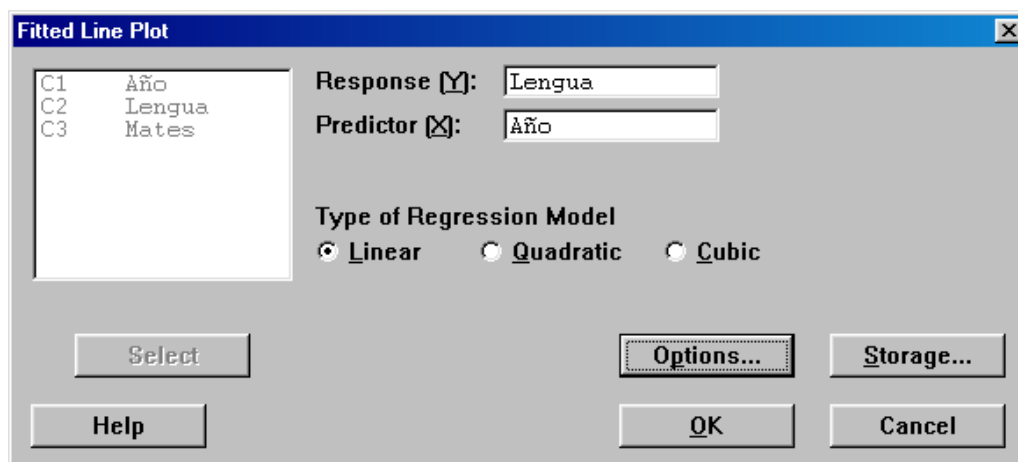
Las medias de las puntuaciones obtenidas en cada test se muestran a continuación (las puntuaciones utilizan una escala distinta a la decimal para evitar ser interpretadas fuera del ámbito del estudio):

Año	Lengua	Mates
1986	466	492
1987	466	492
1988	463	493
1989	460	488
1990	455	488
1991	453	484
1992	445	481
1993	444	480
1994	434	472
1995	431	472
1996	429	470
1997	429	468
1998	427	467
1999	424	466
2000	424	466

1. Representar la nube de puntos junto con la recta de regresión para cada uno de los pares año-test.

¿Podemos decir que ambas puntuaciones varían a la misma velocidad con el paso de los años?

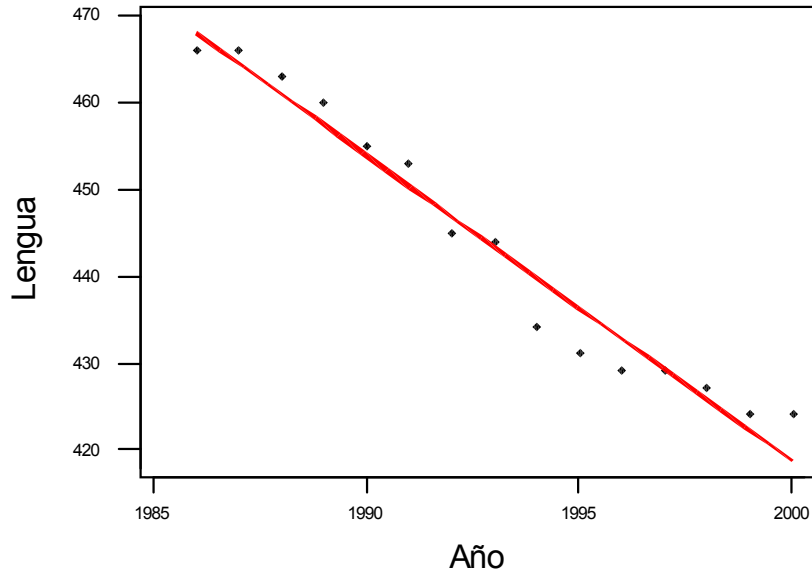
Seleccionamos *Stat > Regression > Fitted Line Plot* :



Hacemos lo propio con la variable Mates. Los gráficos, junto con las correspondientes rectas de regresión, se muestran a continuación:

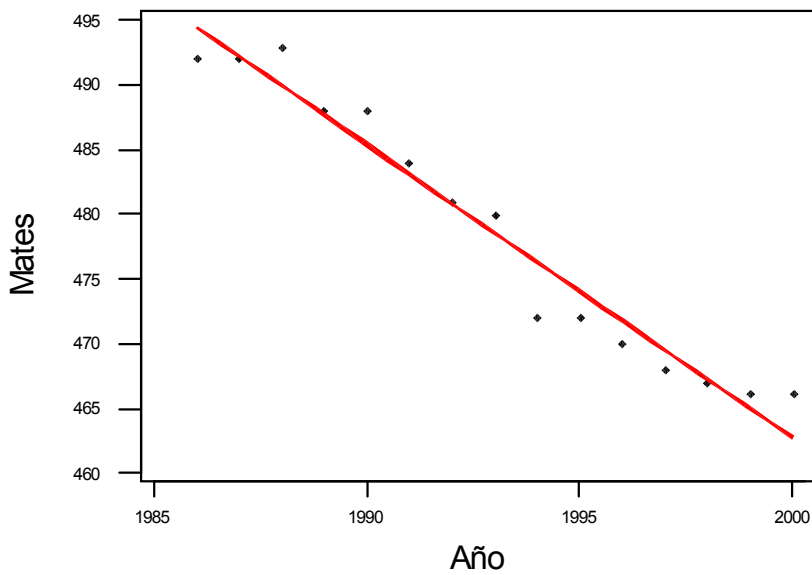
Regression Plot

$$Y = 7440,19 - 3,51071X$$
$$R\text{-Sq} = 96,1 \%$$



Regression Plot

$$Y = 4998,44 - 2,26786X$$
$$R\text{-Sq} = 95,6 \%$$

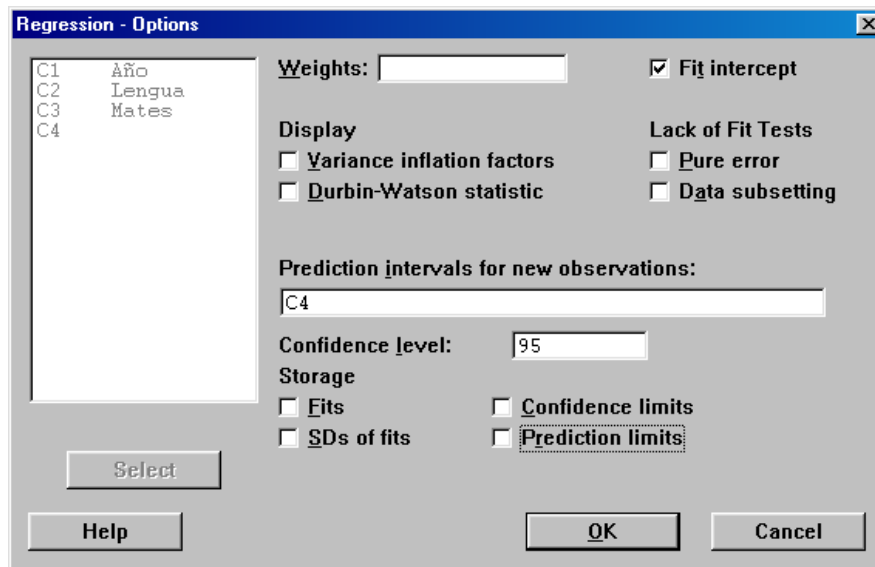
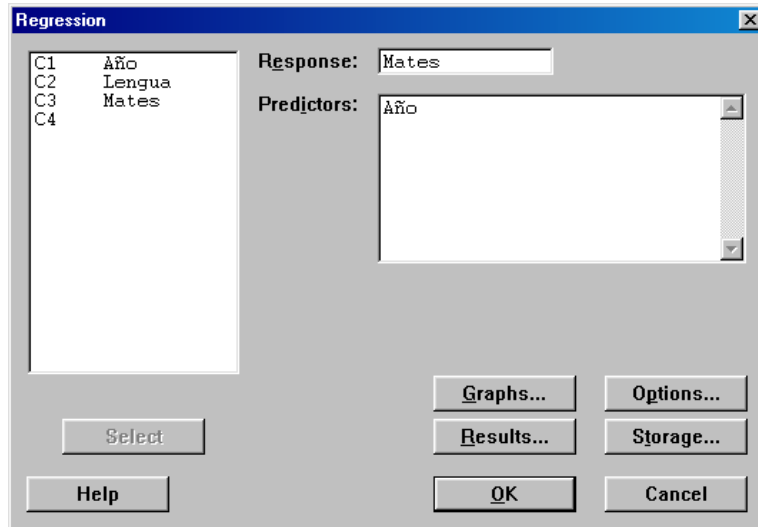


Se aprecia un descenso lineal en ambas puntuaciones conforme pasan los años. Sin embargo, podemos ver que el descenso es más acentuado en las calificaciones de Lengua, ya que aquí la pendiente de la recta es de $-3,51$ por una pendiente de $-2,27$ en el caso de Mates (i.e.: las notas de Lengua parecen decrecer más rápidamente que las de Mates).

2. A partir de los modelos lineales anteriores, pronosticar el valor de la puntuación de Mates para los años 1960, 1990, 2005, y 2010. ¿Cuáles de estos pronósticos tienen sentido?

Colocamos los valores 1960, 1990, 2005 y 2010 (por este orden) en una nueva columna, digamos la C4

Seleccionamos *Stat > Regression > Regression* :



A continuación se muestra parte del output que genera el programa. Bajo la columna *Fit* aparecen las puntuaciones que el modelo predice para cada uno de los años anteriores. El mismo programa nos avisa de que todas excepto la segunda son poco fiables. Ello se debe a que corresponden a años que caen fuera del rango sobre el que disponemos de datos (desde 1986 al 2000).

Regression Analysis

Predicted Values

Fit	StDev Fit	95,0% CI	95,0% PI	
553,439	4,465	(543,793; 563,085)	(542,643; 564,236)	XX
485,404	0,706	(483,879; 486,928)	(480,320; 490,487)	
451,386	1,711	(447,689; 455,082)	(445,288; 457,483)	X
440,046	2,353	(434,963; 445,130)	(433,021; 447,072)	XX

X denotes a row with X values away from the center

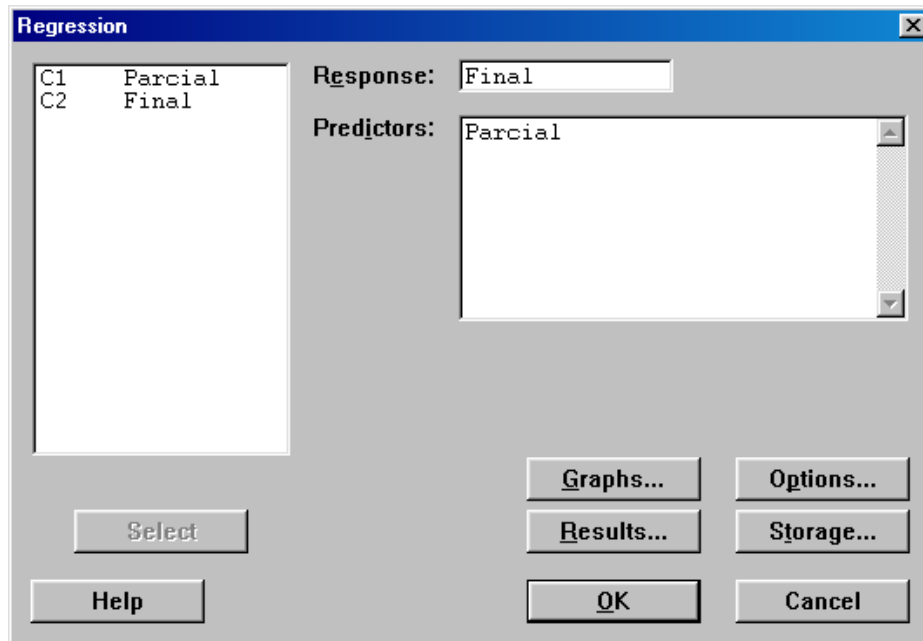
XX denotes a row with very extreme X values

CASO 5-3: EXAMEN PARCIAL Y EXAMEN FINAL

En el archivo **examenes.mtw** están contenidas las notas obtenidas por cada alumno de una asignatura en un examen parcial y en el examen final. Se pide:

1. Construir un modelo lineal para explicar la nota obtenida en el final a partir de la obtenida en el parcial y calcular el intervalo de confianza a nivel del 90% para la pendiente de la recta de regresión.

Seleccionamos *Stat > Regression > Regression* :



El output del programa nos proporciona la recta de regresión: $\text{Final} = 2,25 + 0,755 \text{ Parcial}$. Además, entre otras cosas, podemos ver que este modelo permite explicar aproximadamente un 50% del comportamiento de la variable Final a partir del de la variable Parcial (ver R-Sq).

Regression Analysis

The regression equation is

$$\text{Final} = 2,25 + 0,755 \text{ Parcial}$$

Predictor	Coef	StDev	T	P
Constant	2,247	1,022	2,20	0,036
Parcial	0,7546	0,1417	5,32	0,000

S = 1,151 R-Sq = 49,4% R-Sq(adj) = 47,7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	37,574	37,574	28,35	0,000
Residual Error	29	38,440	1,326		
Total	30	76,014			

Unusual Observations

Obs	Parcial	Final	Fit	StDev Fit	Residual	St Resid
27	7,50	5,200	7,906	0,216	-2,706	-2,39R

R denotes an observation with a large standardized residual

Sabemos que la forma general de un intervalo t-Student de confianza, a nivel $1 - \alpha$, para un determinado parámetro es la siguiente:

$$(\text{parámetro}) \pm t(\alpha/2, n-2) * (\text{Stdev}),$$

donde $t(\alpha/2, n-2)$ es el valor que en una distribución t-Student con $n-2$ grados de libertad deja a su izquierda una probabilidad de $1 - \alpha/2$, siendo n el número de observaciones.

En este caso,

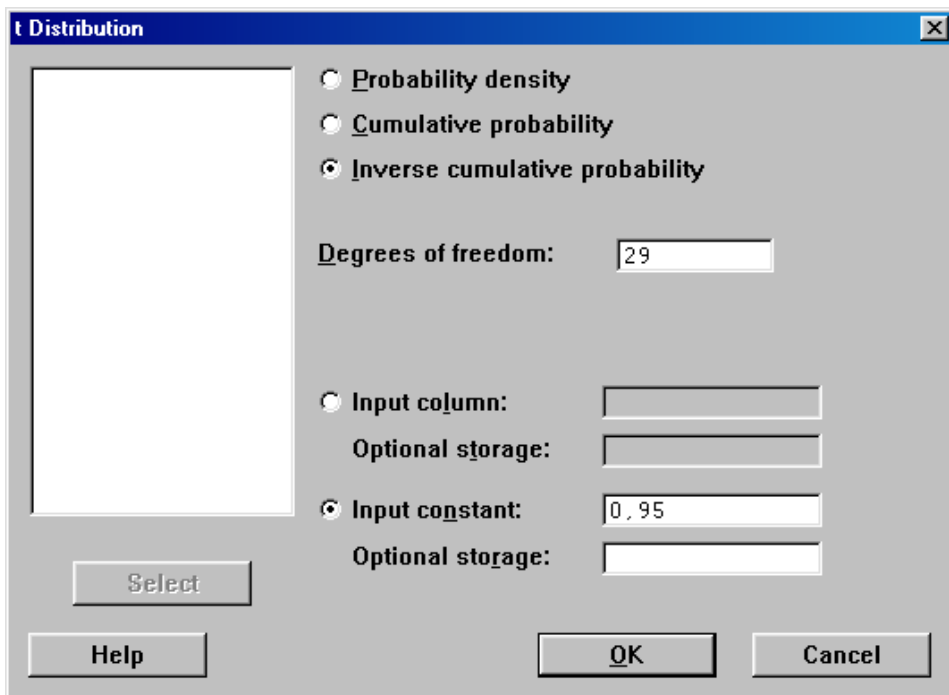
$$1 - \alpha = 0,90 \rightarrow \alpha = 0,10 \rightarrow 1 - \alpha/2 = 0,95$$

$$n = 31 - 2 = 29$$

$$\text{parámetro} = 0,7546$$

$$\text{Stdev} = 0,1417$$

Seleccionamos **Calc > Probability Distributions > t** :



Inverse Cumulative Distribution Function

Student's t distribution with 29 DF

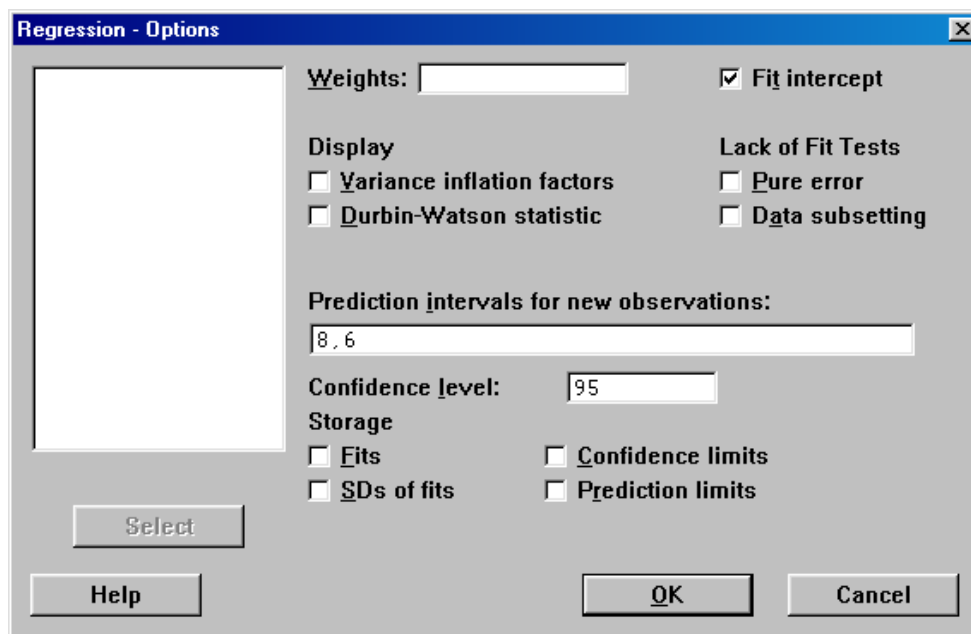
P (X <= x)	x
0,9500	1,6991

Tendremos pues que el intervalo deseado es $(0,7546) \pm (1,6991)*(0,1417) = (0,5138, 0,9954)$

2. Hallar el intervalo de confianza a nivel del 95% para la media de las notas finales correspondientes a todas aquellas personas que obtuvieron una puntuación de 8,6 en el parcial. Sabiendo que una persona ha obtenido una puntuación de 8,6 en el parcial, hallar un intervalo de predicción a nivel del 90% para el valor que se espera obtenga en el final.

Observemos que nos están pidiendo dos cosas diferentes: por un lado nos hablan de hallar un **intervalo de confianza para la media** de las notas en el final correspondientes a **todas** aquellas personas que han obtenido una determinada nota en el parcial, mientras que por otro nos piden un **intervalo de predicción para la nota** que se espera obtenga en el final una **única** persona que ha obtenido una determinada nota parcial. Como siempre resulta más fácil hacer predicciones sobre medias que sobre parámetros individuales, es de esperar que el intervalo de predicción contenga (sea mayor) que el intervalo de confianza.

Seleccionamos *Stat > Regression > Regression > Options* :

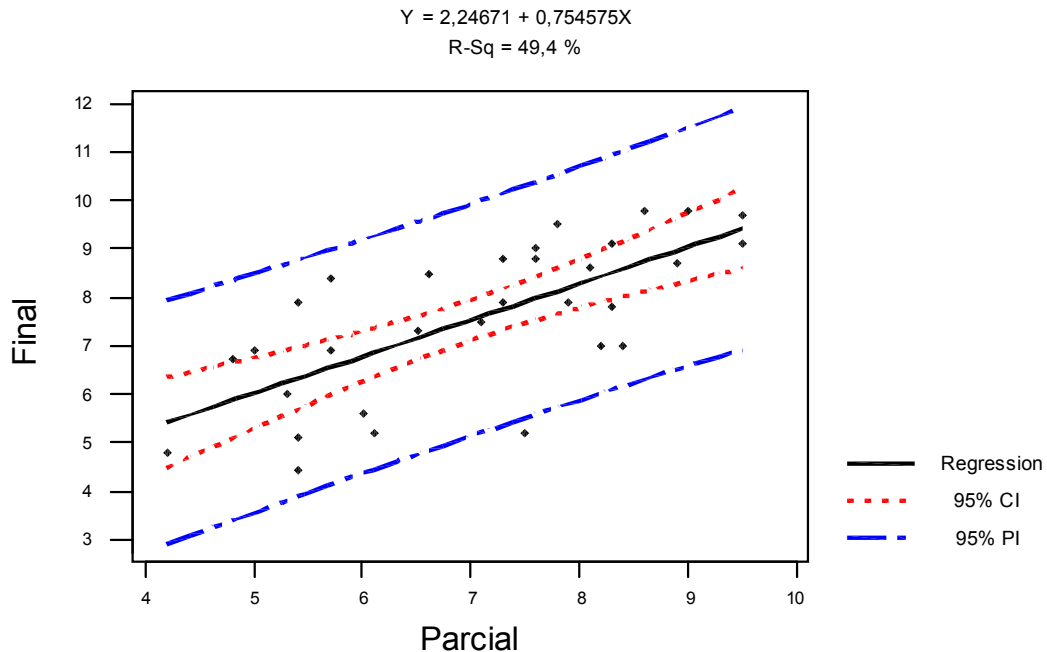


Predicted Values				
Fit	StDev Fit	95,0% CI		95,0% PI
8,736	0,300	(8,122;	9,350)	(6,303; 11,169)

En este caso, el modelo construido predice que una persona que hubiese obtenido una puntuación de 8,6 en el parcial obtendría una puntuación de 8,736 en el final. La nota final media de todos aquellos que obtuvieron en el parcial un 8,6 estará, con probabilidad 0,95, entre 8,12 y 9,35. Por otra parte, si una persona ha obtenido un 8,6 en el parcial, es de esperar (con probabilidad 0,95) que obtenga en el final una nota situada entre 6,30 y 11,17.

Observar que, tal y como predijimos, el intervalo de confianza para la media está contenido dentro del intervalo de predicción para una observación individual. Este hecho se observa mejor en el siguiente gráfico, obtenido a partir de las opciones *Stat > Regression > Fitted Line Plot > Options* :

INTERVALOS DE CONFIANZA E INTERVALOS DE PREDICCIÓN



3. **Contrastar, para un nivel de significación $\alpha = 0,05$, la hipótesis nula H_0 : el coeficiente de la variable Parcial (variable X) es cero (hipótesis que es equivalente a H_0 : el coeficiente de correlación lineal de la población, ρ , es cero). En otras palabras, al hacer esta hipótesis nos estamos preguntando si hay indicios suficientes o no para considerar que el modelo obtenido es válido (la hipótesis nula afirma que no lo es).**

Para responder a la pregunta de si el modelo construido es o no válido para explicar el comportamiento de la variable Y en función del de la variable X, basta con volver al output inicial:

Regression Analysis					
The regression equation is					
Final = 2,25 + 0,755 Parcial					
Predictor	Coef	StDev	T	P	
Constant	2,247	1,022	2,20	0,036	
Parcial	0,7546	0,1417	5,32	0,000	
S = 1,151		R-Sq = 49,4%		R-Sq(adj) = 47,7%	
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	37,574	37,574	28,35	0,000
Residual Error	29	38,440	1,326		
Total	30	76,014			

Observar que el output nos proporciona ya el p-valor asociado al test cuya hipótesis nula afirma que el coeficiente de la X es cero (el modelo no sirve). Dicho p-valor es 0,000, por lo cual rechazaremos la hipótesis nula y concluiremos que el modelo sí es válido. En la parte de Analysis of Variance se realiza el test equivalente sobre el coeficiente poblacional de correlación lineal, por lo que no es de extrañar que obtengamos el mismo p-valor. Finalmente, notar que también se realiza un test similar sobre el término independiente del modelo (cuyo p-valor en este caso es de 0,036).

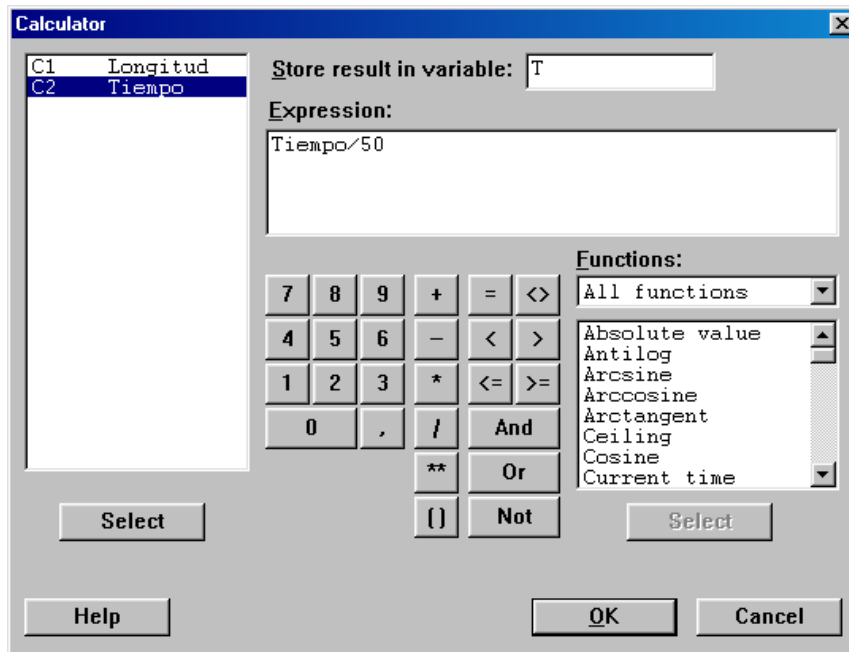
CASO 5-4: MOVIMIENTO PENDULAR

En una clase de física se lleva a cabo un experimento consistente en dejar caer un péndulo de longitud variable desde su posición horizontal, esperar a que complete 50 ciclos y registrar el tiempo transcurrido. Los datos obtenidos en 5 pruebas se muestran a continuación:

	Longitud	Tiempo
1	175,2	132,5
2	151,5	123,4
3	126,4	112,8
4	101,7	101,2
5	77,0	88,2

1. Sea T el tiempo medio por ciclo. Calcular T para cada prueba dividiendo el tiempo transcurrido por 50. Representar T vs. Longitud y hallar la recta de regresión que permite explicar T a partir de la Longitud. ¿Te parece bueno el modelo hallado?

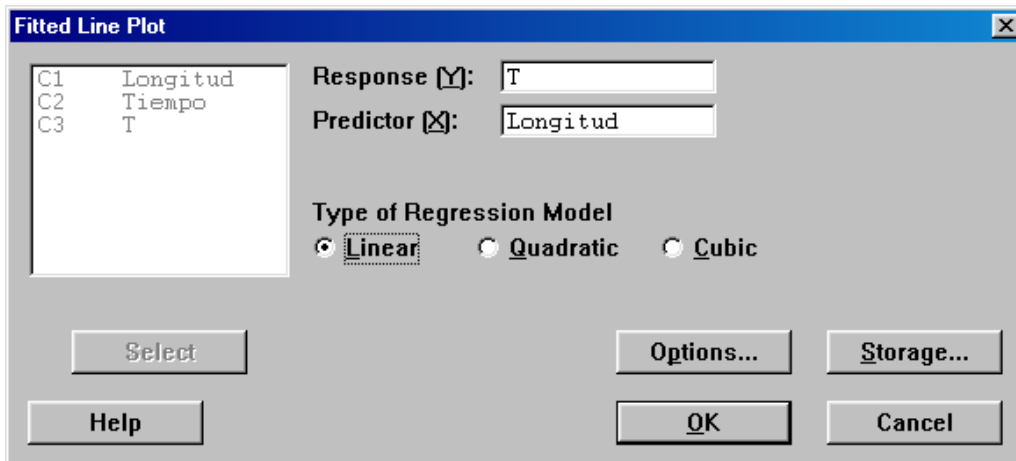
Seleccionamos *Calc > Calculator* :



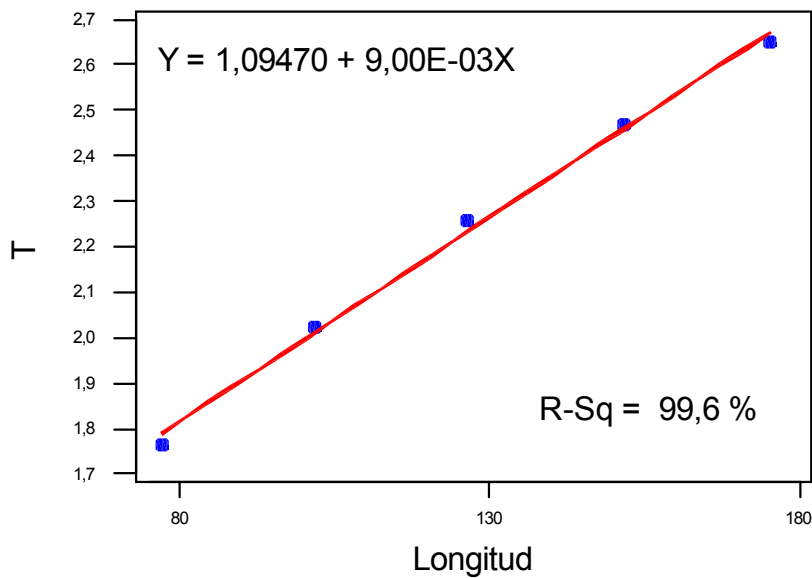
Con ello generaremos una nueva columna que contendrá los valores de T :

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
↓	Longitud	Tiempo	T								
1	175,2	132,5	2,650								
2	151,5	123,4	2,468								
3	126,4	112,8	2,256								
4	101,7	101,2	2,024								
5	77,0	88,2	1,764								
6											
7											
8											

Seleccionamos *Stat > Regression > Fitted Line Plot* :



Regression Plot

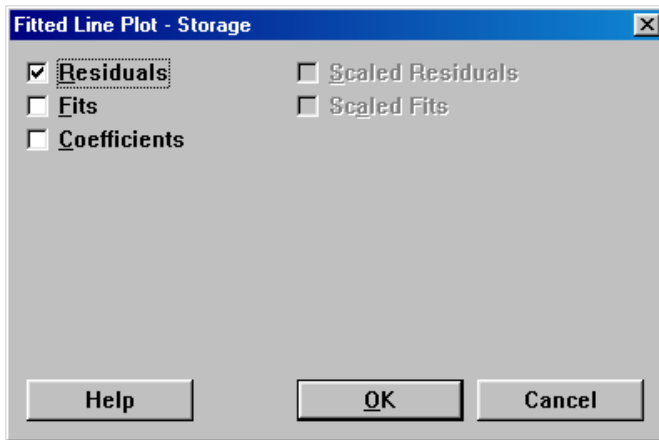


Tanto la gráfica de regresión como el valor del coeficiente de determinación $R-sq = 0,996$ parecen indicar que el modelo hallado es bastante útil a la hora de estimar T a partir de la variable Longitud: los puntos se aproximan bastante a la recta de regresión y, además, el coeficiente de determinación nos dice que aproximadamente un 99,6% de la variación en T es explicable con este modelo a partir del comportamiento de la variable Longitud.

Sin embargo, como veremos en el siguiente apartado, sería prematuro confiar en este modelo antes de analizar si se cumplen las hipótesis de la regresión lineal.

2. Representar en un gráfico los residuos del modelo anterior vs. la Longitud. ¿Hay alguna indicación que os haga pensar en la necesidad de usar otro modelo distinto del lineal?

- ☐ Seleccionamos la subopción **Storage** de la ventana anterior y pedimos al programa que nos guarde los residuos en una nueva columna:



- ☐ Ahora vamos a **Graph > Plot** :

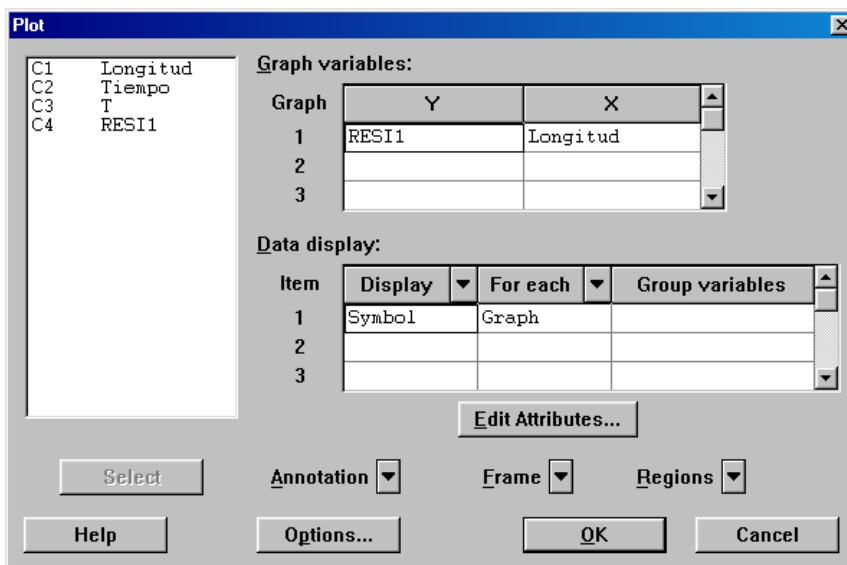
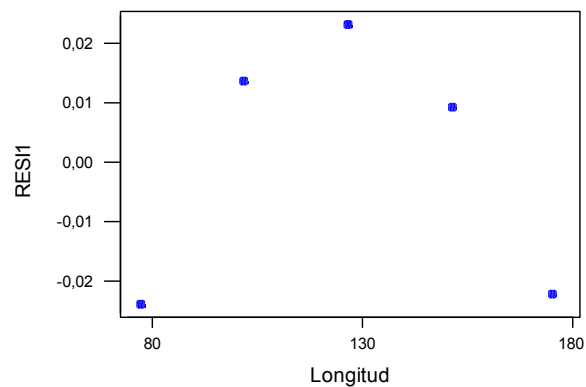
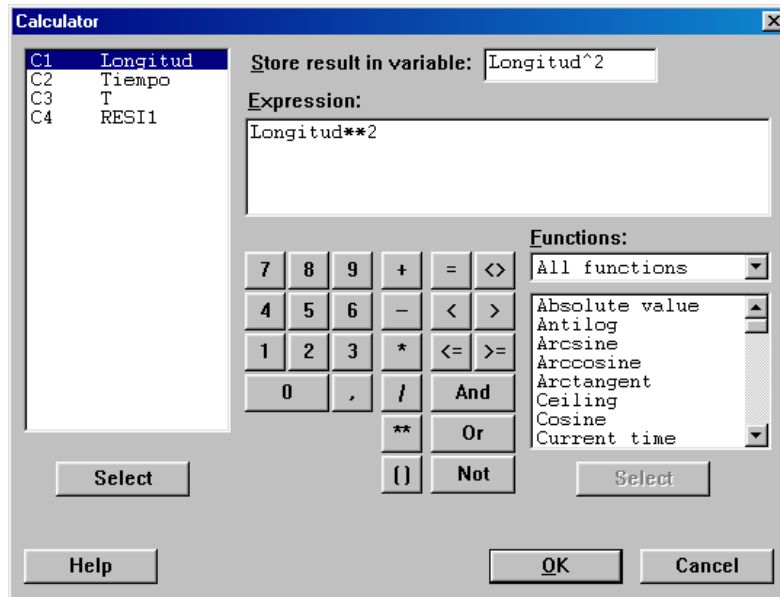


GRÁFICO DE RESÍDUOS VS. LONGITUD

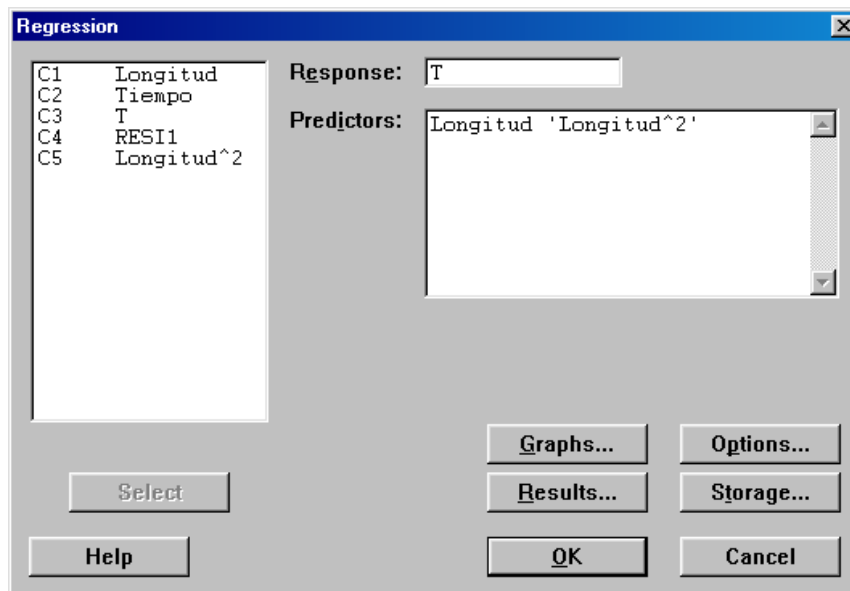


Observamos en el gráfico de residuos anterior una clara tendencia parabólica, lo que implica que el modelo anterior no es válido. Debemos buscar un modelo que se ajuste mejor a los datos obtenidos con nuestro experimento. Vamos a probar con un modelo polinómico de orden 2 (notar que éste ya no será un modelo lineal). Usaremos las variables Longitud y Longitud^2 (el cuadrado de la variable anterior):

☞ Seleccionamos **Calc > Calculator** :



☞ Seleccionamos **Stat > Regression > Regression** :



Regression Analysis

The regression equation is

$$T = 0,812 + 0,0139 \text{ Longitud} - 0,000019 \text{ Longitud}^2$$

S = 0,001945 R-Sq = 100,0% R-Sq(adj) = 100,0%

¡Vaya! Ahora sí parece que hemos dado con un buen modelo. Observar que el coeficiente de determinación es 1, lo que significa que con este modelo podemos explicar de forma total el comportamiento de T a partir del de la variable Longitud.