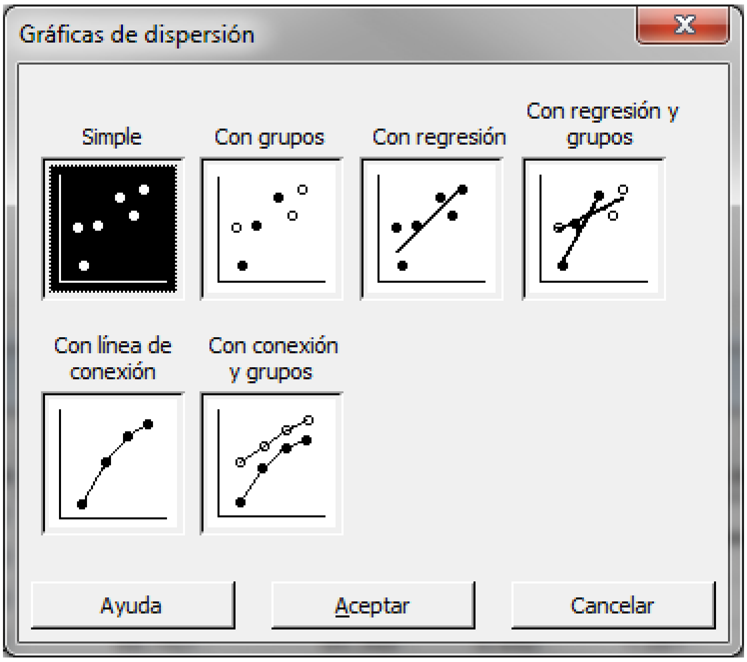
Regresión lineal simple. Prof. Dr. Víctor Yepes

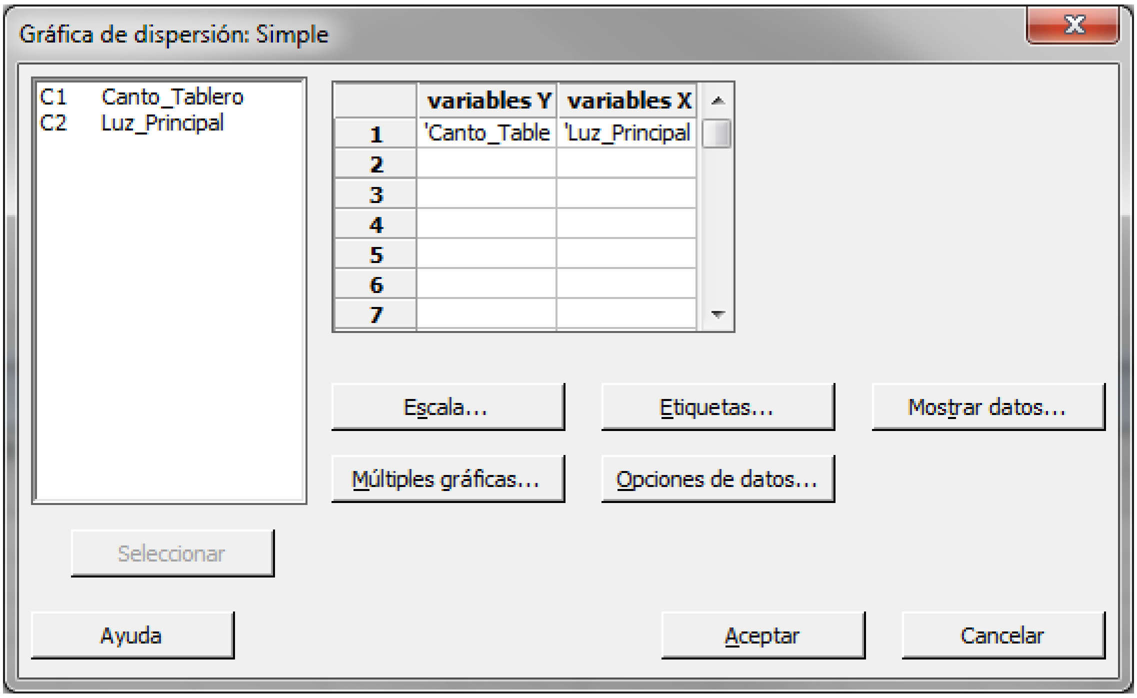
Se dispone de una muestra de 26 puentes losa postesados macizos. Los resultados se muestran en el cuadro siguiente. Se quiere conocer la relación que existe entre la luz principal de este tipo de puentes y el canto del tablero. Utilizaremos los programas siguientes: MINITAB, SPSS, EXCEL Y MATLAB.

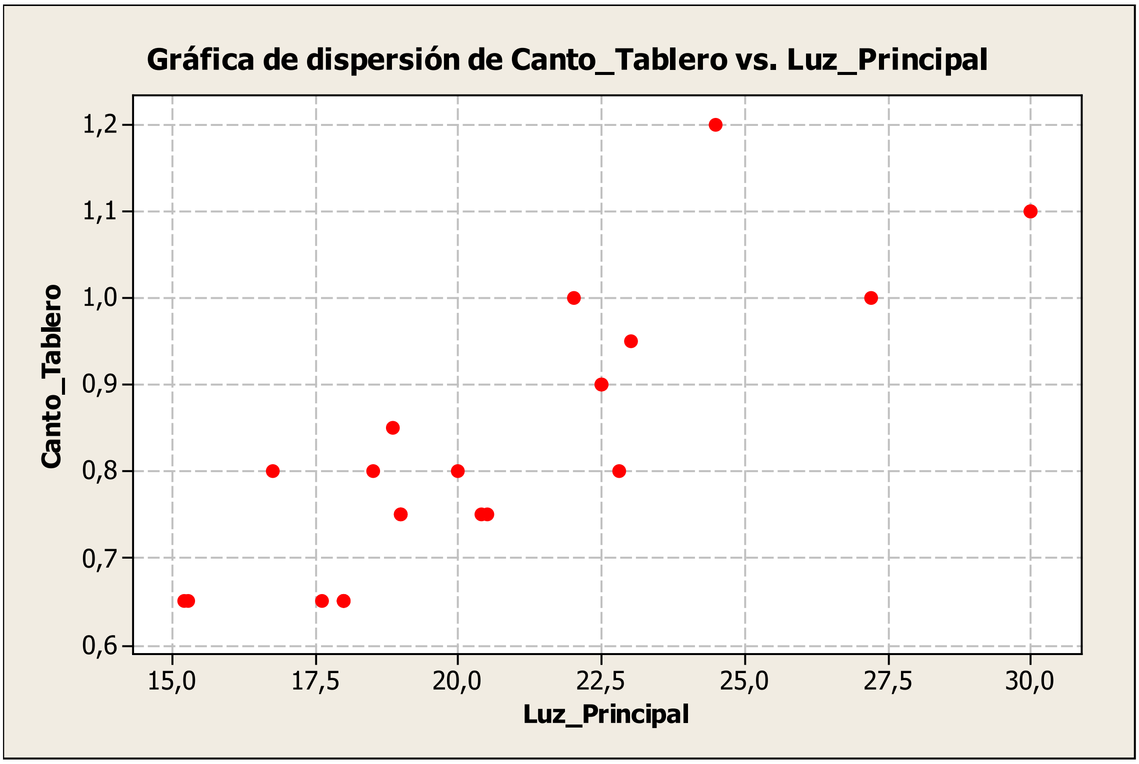
Número	Luz principal (m)	Canto del tablero (m)
1	30,00	1,10
2	30,00	1,10
3	30,00	1,10
4	30,00	1,10
5	22,50	0,90
6	19,00	0,75
7	22,50	0,90
8	18,00	0,65
9	18,00	0,65
10	17,60	0,65
11	20,00	0,80
12	24,50	1,20
13	19,00	0,75

Número	Luz principal (m)	Canto del tablero (m)
14	15,27	0,65
15	15,20	0,65
16	18,00	0,65
17	20,40	0,75
18	27,20	1,00
19	20,50	0,75
20	22,80	0,80
21	18,85	0,85
22	16,75	0,80
23	18,50	0,80
24	22,50	0,90
25	22,00	1,00
26	23,00	0,95

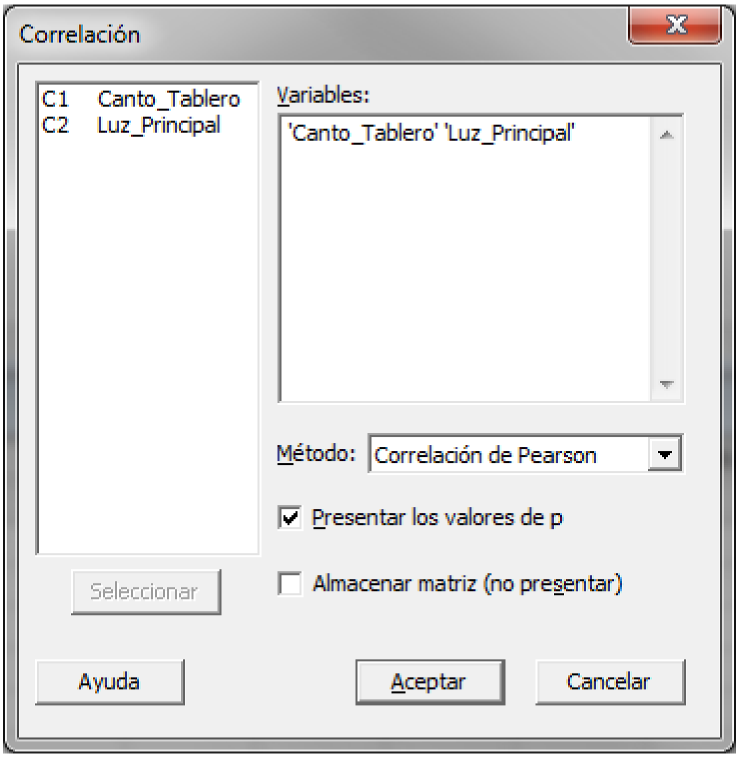
Vamos a analizar con Minitab. Empezaremos por representar la gráfica de dispersión de los datos para comprobar cómo se comportan desde el punto de vista cualitativo. Gráfica > Gráfica de dispersión simple




Gráfica de dispersión: Simple
C1 Canto_Tablero
C2 Luz_Principal
variables Y
variables X
1 'Canto_Table 'Luz_Principal
2
3
4
5
6
7
Escala...
Etiquetas...
Mostrar datos...
Múltiples gráficas...
Opciones de datos...
Seleccionar
Ayuda
Aceptar
Cancelar


Gráfica de dispersión de Canto_Tablero vs. Luz_Principal
Canto_Tablero
1,2
1,1
1,0
0,9
0,8
0,7
0,6
15,0
17,5
20,0
22,5
25,0
27,5
30,0
Luz_Principal

Ahora calcularemos e interpretaremos el coeficiente de correlación entre estas dos variables. Estadística > Estadística básica > Correlación



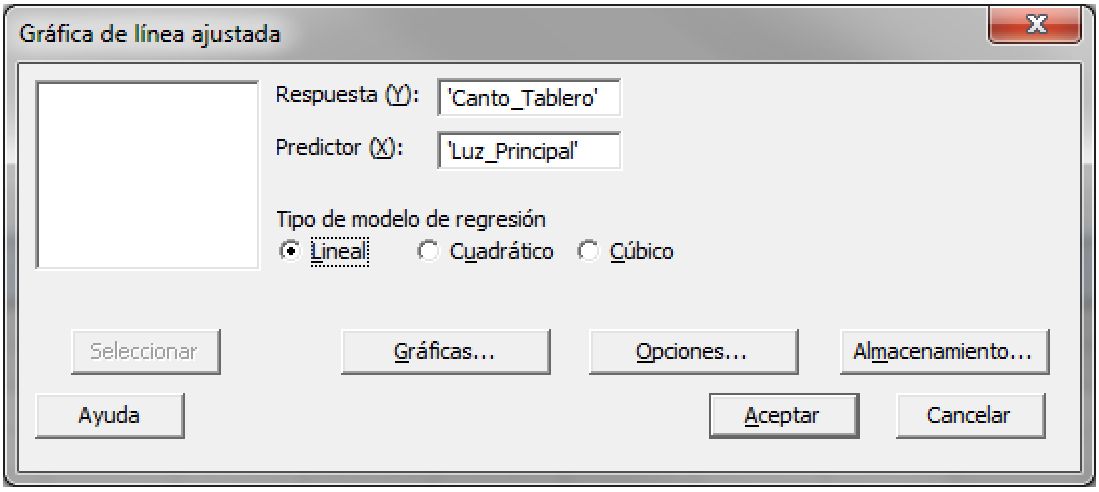
El resultado es el siguiente:

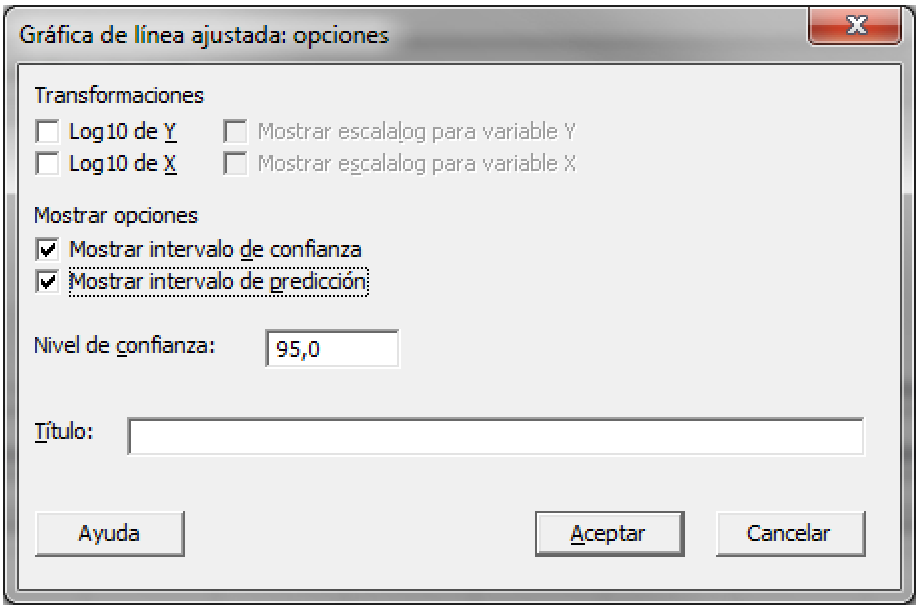
Correlaciones: Luz_Principal; Canto_Tablero

```
Correlación de Pearson de Luz_Principal y Canto_Tablero = 0,886
Valor P = 0,000
```

El coeficiente de correlación de Pearson es una medida de la relación lineal entre dos variables aleatorias cuantitativas, que es independiente de la escala de medida de las variables. Aquí se observa que el coeficiente de correlación de Pearson supera el 80%, lo cual indica una buena relación entre las variables. Además, como p-valor < 0.05, se comprueba que el coeficiente es altamente significativo.

Vamos ahora a ajustar una recta a la nube de puntos por mínimos cuadrados, con un nivel de confianza del 95%. Estadísticas > Regresión > Gráfica de línea ajustada > Opciones: Mostrar intervalo de confianza; Mostrar intervalo de predicción



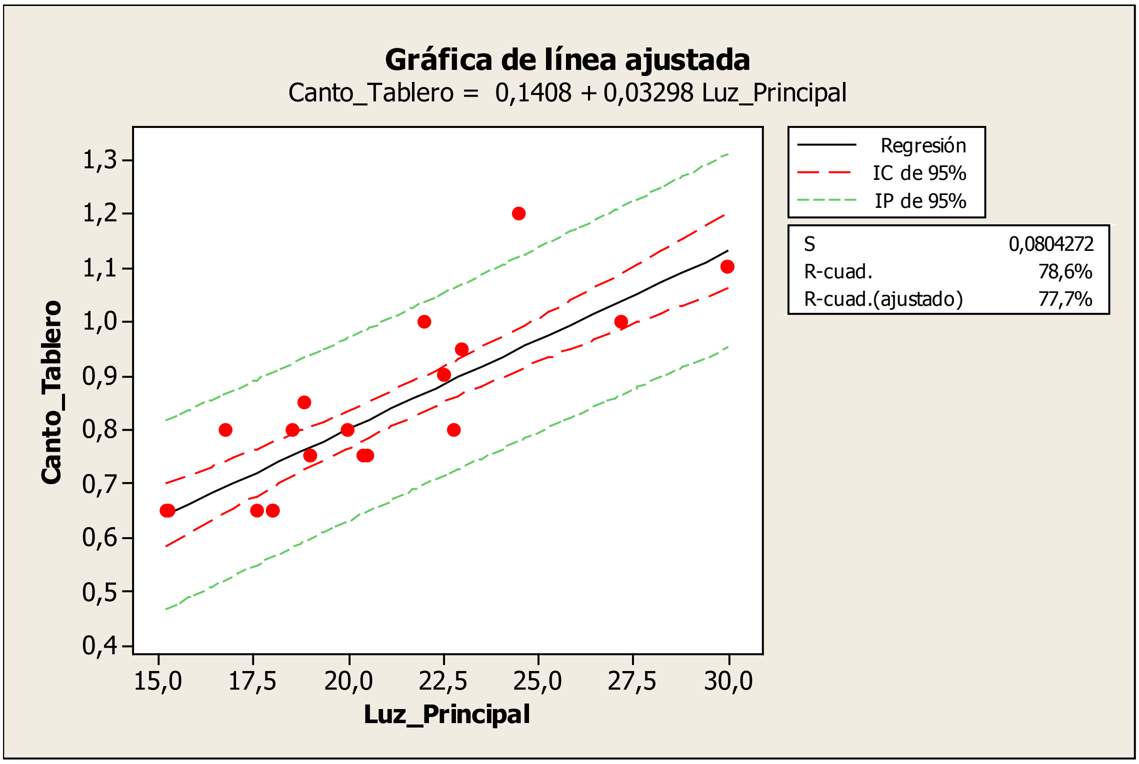


Análisis de regresión: Canto_Tablero vs. Luz_Principal

```
La ecuación de regresión es
Canto_Tablero = 0,1408 + 0,03298 Luz_Principal


S = 0,0804272   R-cuad. = 78,6%   R-cuad.(ajustado) = 77,7%


Análisis de varianza

Fuente      GL        SC        MC      F      P
Regresión    1  0,569370  0,569370  88,02  0,000
Error       24  0,155245  0,006469
Total       25  0,724615
```

El R^2 indica que el poder de explicación del modelo de mínimos cuadrados es del 77,7%. Es decir, ese es el porcentaje de la variabilidad explicada por el modelo.

La prueba ANOVA es para comprobar si es significativo el modelo propuesto. Se rechazar la hipótesis nula de la independencia lineal entre *x* e *y*.



La notación de la gráfica es la siguiente:

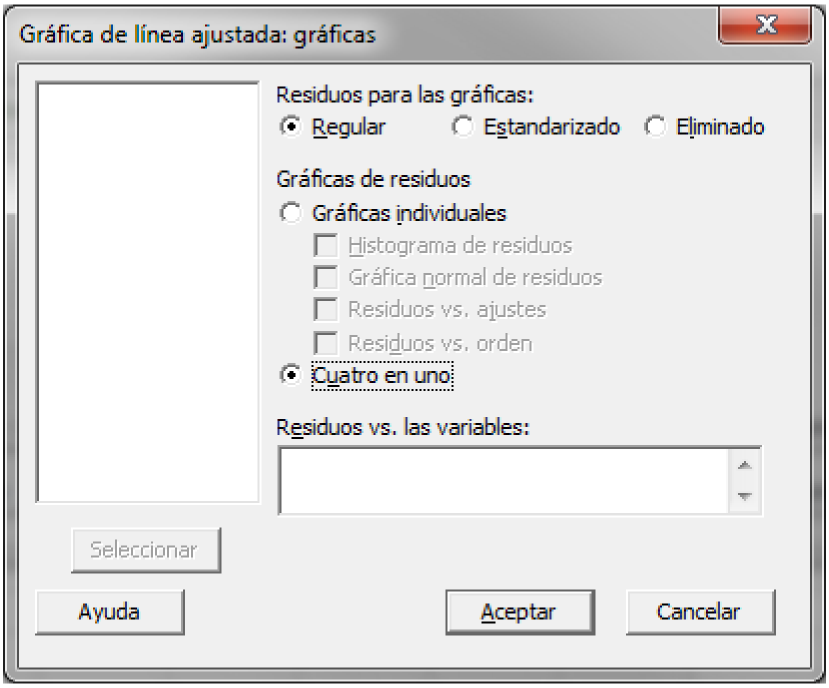
S: Desviación tipo de los residuos (residuo = valor real de la respuesta – valor previsto usando la ecuación).

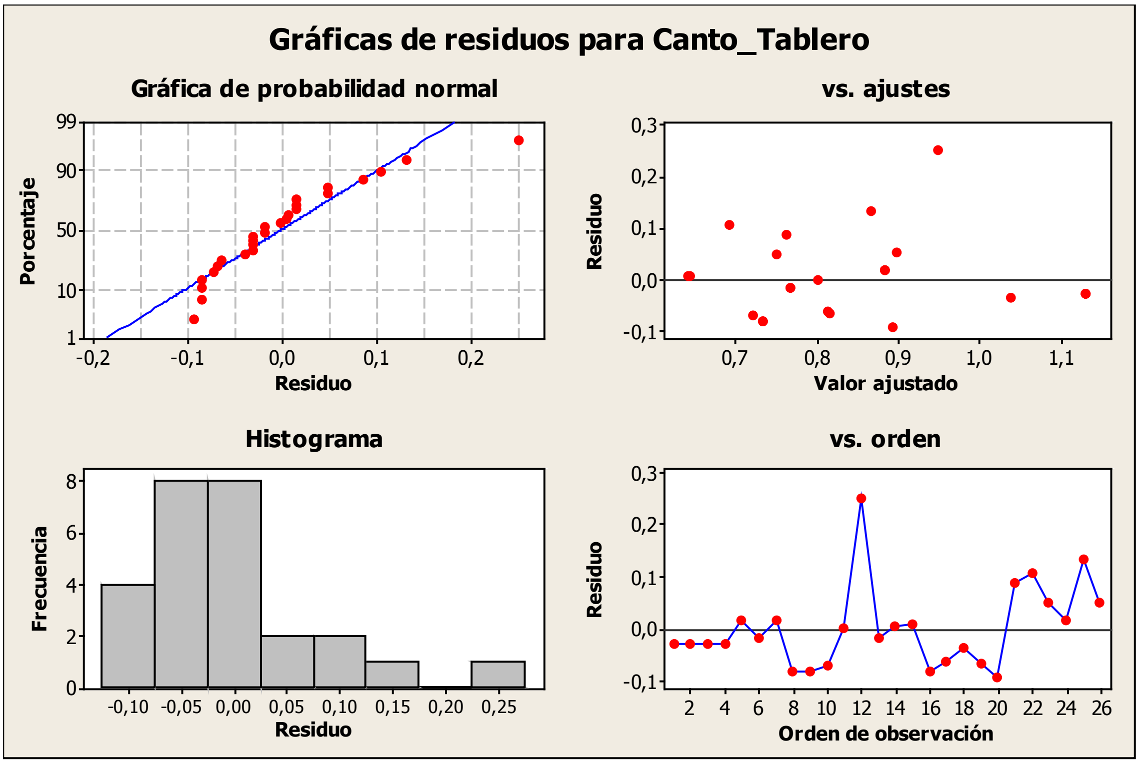
R-cuad. Coeficiente de determinación (R^2): medida de calidad del ajuste. Es el cuadrado del coeficiente de correlación (x100 porque lo da en %).

R-cuad. (ajustado) Coeficiente de determinación ajustado. Medida de la calidad del ajuste utilizada en las ecuaciones de regresión.

IC de 95%: Intervalo de confianza para el valor medio de Y dado un valor de X (banda más cercana a la recta).

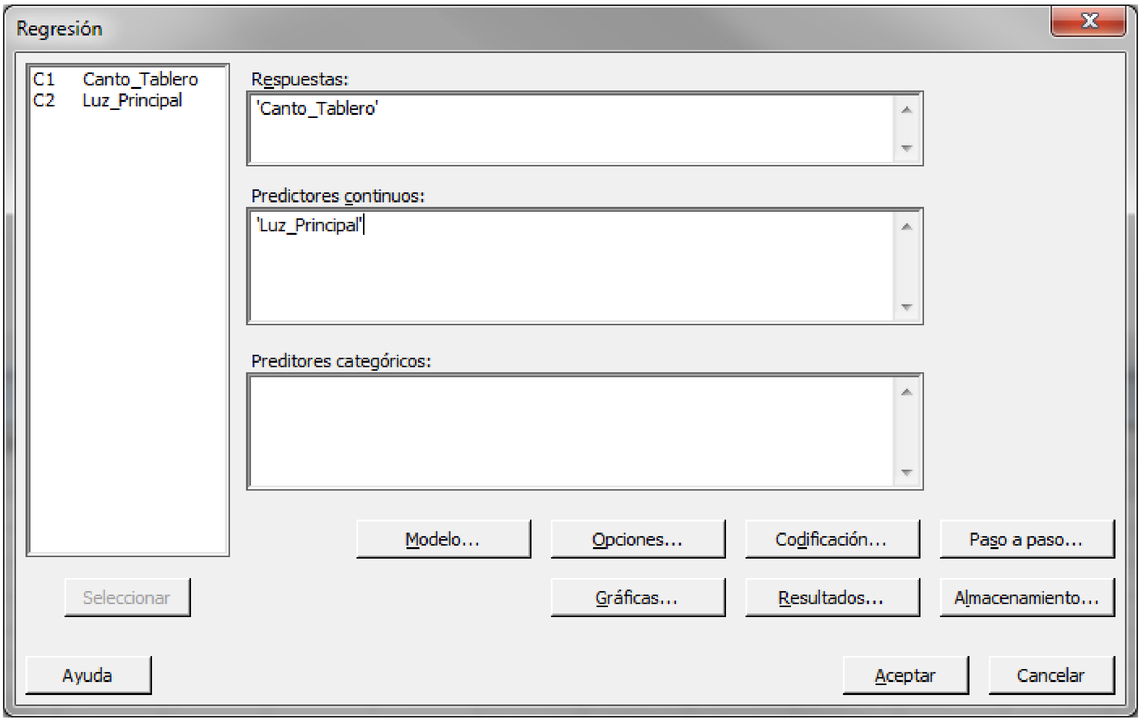
IP de 95%: Intervalo de confianza para los valores individuales de Y dado un valor de X (banda más lejana a la recta).





En la gráfica vemos que los residuos aparecen de forma aleatoria respecto a 0,0, además los residuos se ajustan razonablemente bien a la recta en la gráfica de probabilidad normal, por lo que se cumple la normalidad de los residuos y su varianza constante.

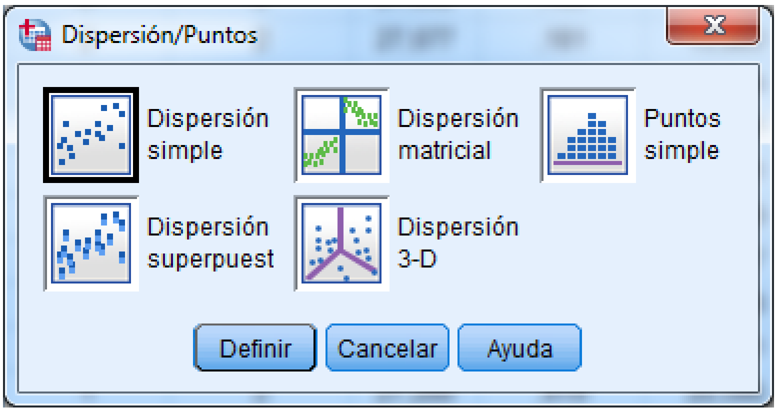
Podemos ahora ver si algún dato se puede considerar como poco común (residuo estandarizado > 2, se marca con R). En otros casos se marcan con X los puntos que tienen mucha influencia sobre la recta. En estos casos se debe valorar la conveniencia de mantener dichos puntos en el estudio. Estadísticas > Regresión > Regresión > Ajustar modelo de regresión




```
Observaciones poco comunes

                                              EE de            Residuo
Obs  Luz_Principal  Canto_Tablero  Ajuste  ajuste  Residuo  estándar
 12           24,5         1,2000  0,9489  0,0187   0,2511     3,21R

R denota una observación con un residuo estandarizado grande.
```

No existen puntos marcados con X, por lo que consideramos todos en el estudio.

Vamos a realizar el estudio con SPSS. En primer lugar, representemos la dispersión de los puntos.
Gráficos > Cuadros de diálogo antiguos > Dispersión/Puntos > Dispersión simple



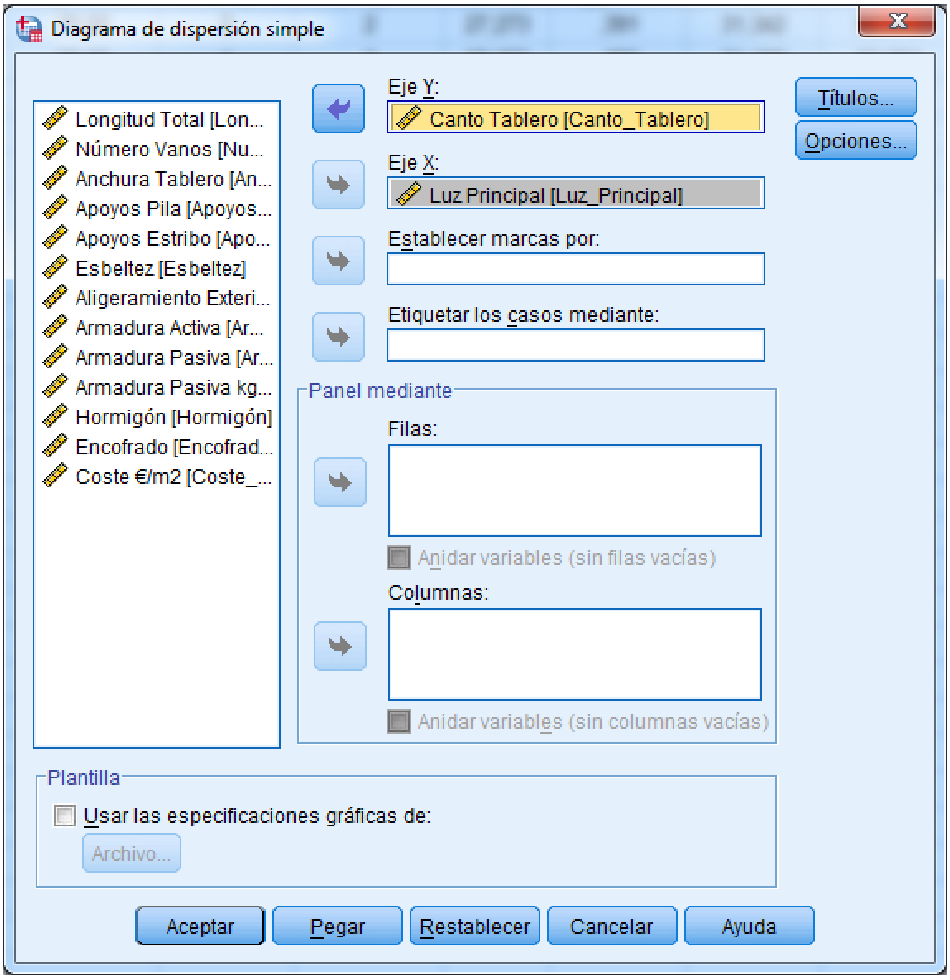
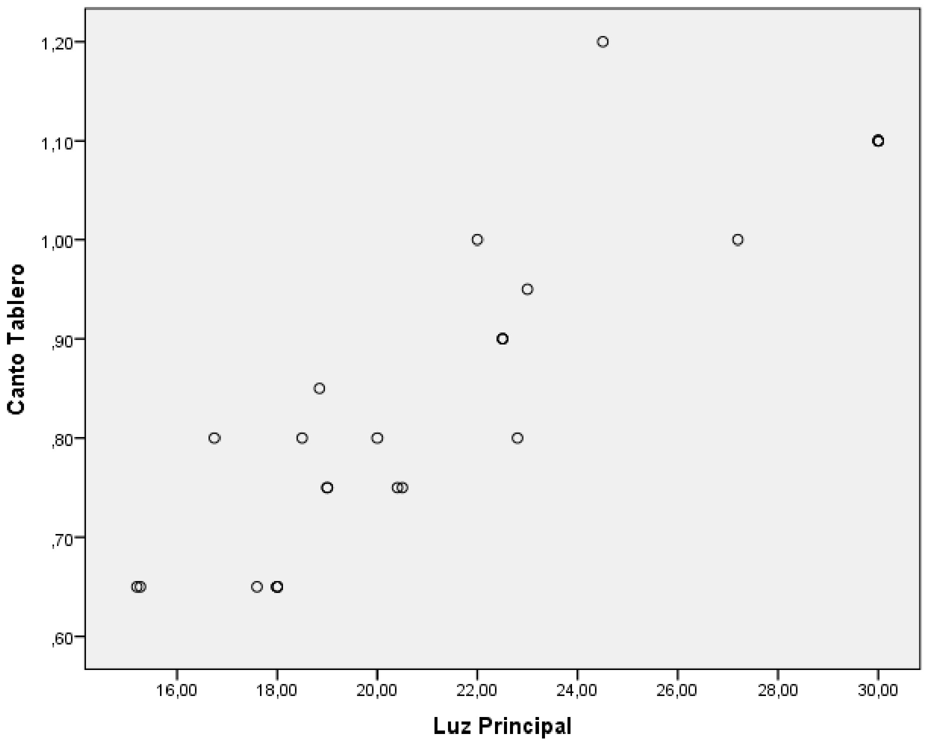
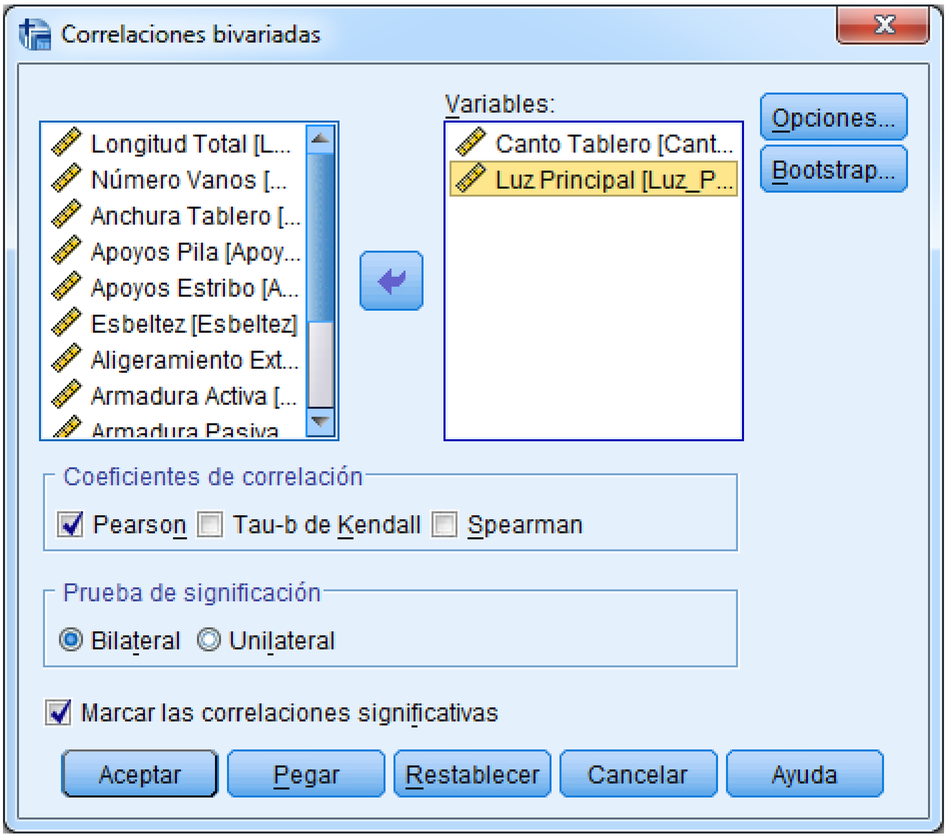


Diagrama de dispersión simple
Longitud Total [Lon...
Número Vanos [Nu...
Anchura Tablero [An...
Apoyos Pila [Apoyos...
Apoyos Estribo [Apo...
Esbeltez [Esbeltez]
Aligeramiento Exteri...
Armadura Activa [Ar...
Armadura Pasiva [Ar...
Armadura Pasiva kg...
Hormigón [Hormigón]
Encofrado [Encofrad...
Coste €/m2 [Coste_...
Eje Y:
Canto Tablero [Canto_Tablero]
Eje X:
Luz Principal [Luz_Principal]
Establecer marcas por:
Etiquetar los casos mediante:
Panel mediante
Filas:
Anidar variables (sin filas vacías)
Columnas:
Anidar variables (sin columnas vacías)
Títulos...
Opciones...
Plantilla
Usar las especificaciones gráficas de:
Archivo...
Aceptar
Pegar
Restablecer
Cancelar
Ayuda


1,20
1,10
1,00
,90
,80
,70
,60
Canto Tablero
16,00
18,00
20,00
22,00
24,00
26,00
28,00
30,00
Luz Principal

Para analizar las correlaciones: Analizar > Correlaciones > Bivariadas > Coeficientes de correlación: Pearson; Tau-b de Kendall; Spearman > Prueba de significación bilateral > Marcar las correlaciones significativas

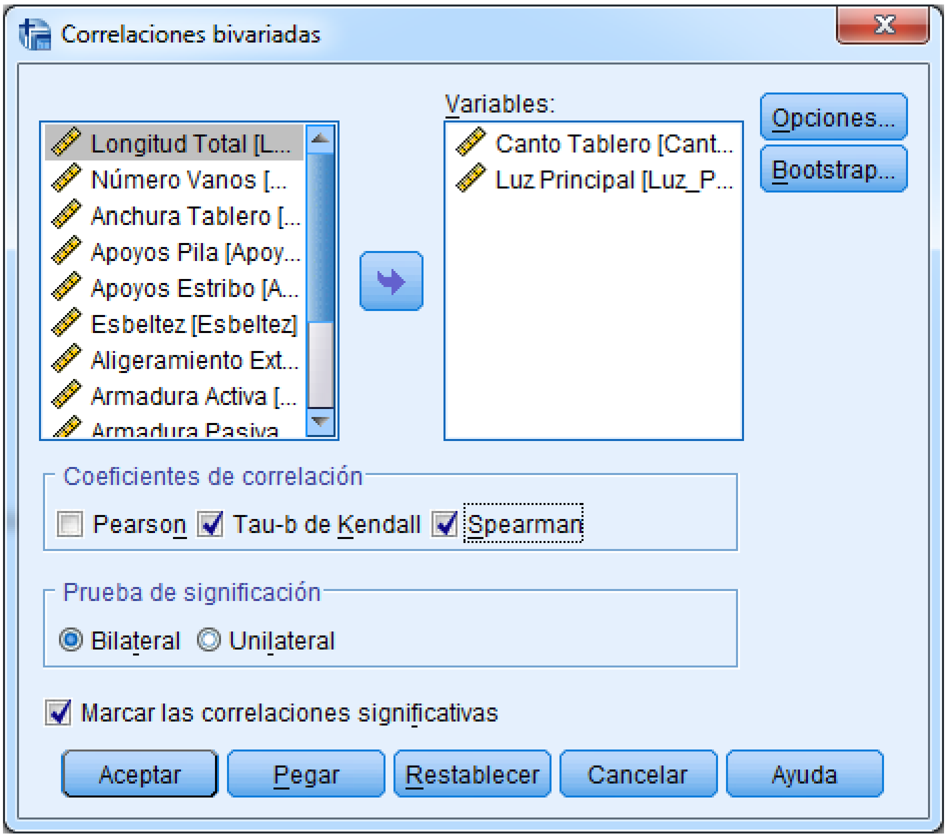


Correlaciones

		Canto Tablero	Luz Principal
Canto Tablero	Correlación de Pearson	1,000	,886**
	Sig. (bilateral)		,000
	N	26	26
Luz Principal	Correlación de Pearson	,886**	1,000
	Sig. (bilateral)	,000	
	N	26	26

**. La correlación es significativa al nivel 0,01 (bilateral).

También se pueden ver las correlaciones no paramétricas (en el caso de no normalidad):

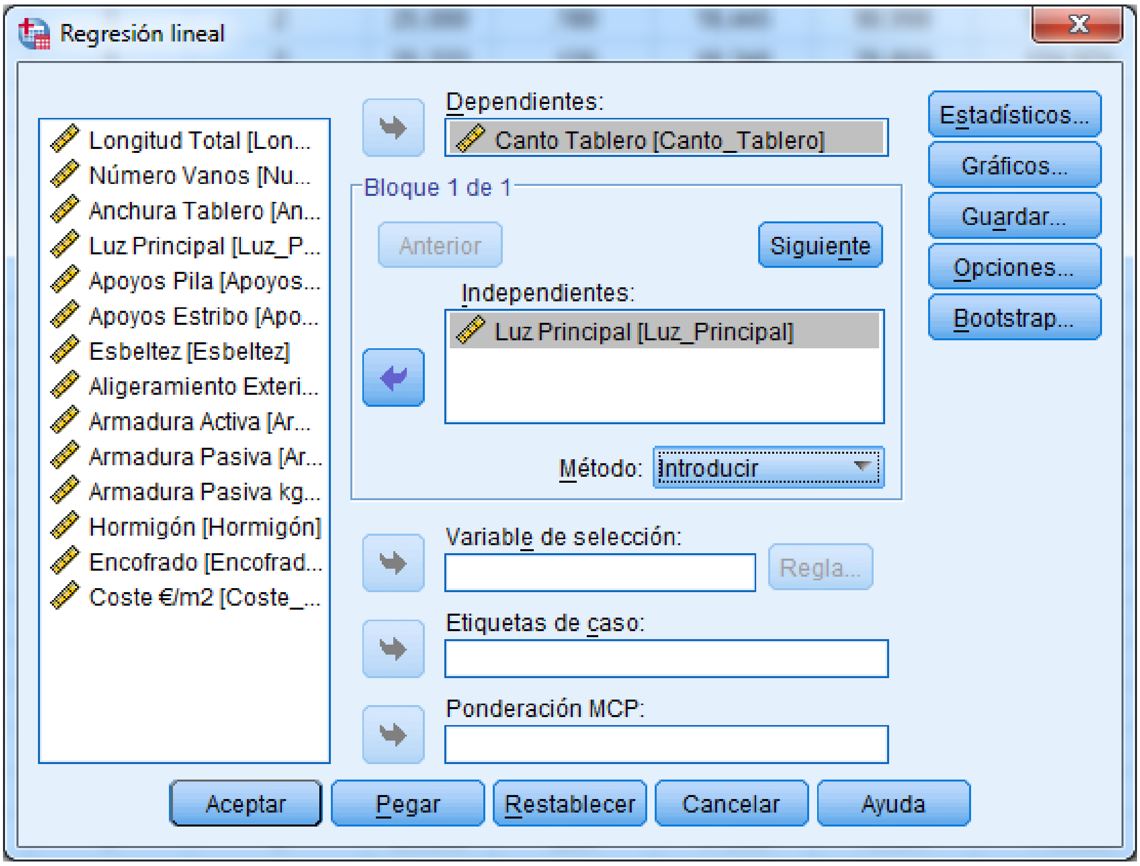


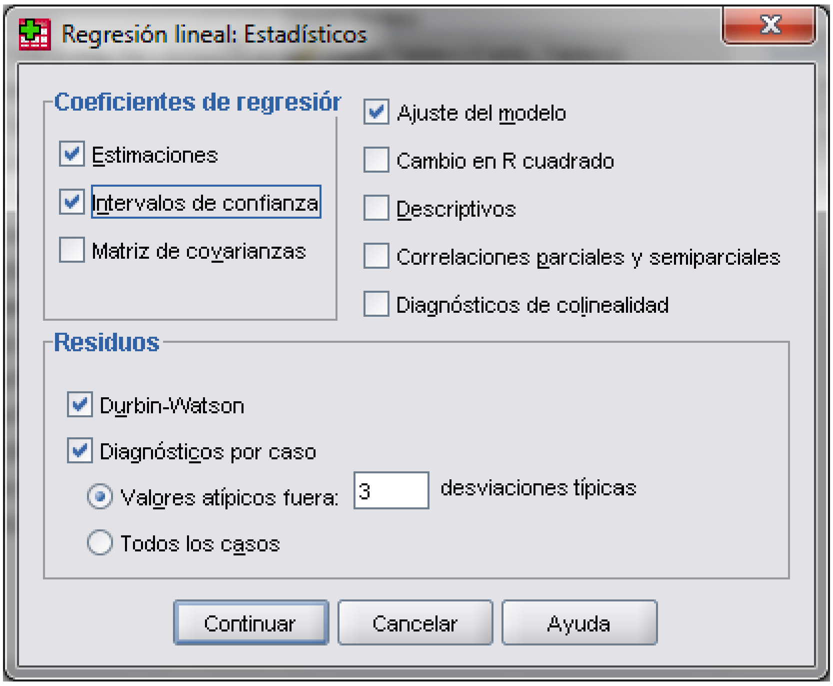
Correlaciones

			Canto Tablero	Luz Principal
Tau_b de Kendall	Canto Tablero	Coeficiente de correlación	1,000	,741**
		Sig. (bilateral)	.	,000
		N	26	26
	Luz Principal	Coeficiente de correlación	,741**	1,000
		Sig. (bilateral)	,000	.
		N	26	26
Rho de Spearman	Canto Tablero	Coeficiente de correlación	1,000	,883**
		Sig. (bilateral)	.	,000
		N	26	26
	Luz Principal	Coeficiente de correlación	,883**	1,000
		Sig. (bilateral)	,000	.
		N	26	26

**. La correlación es significativa al nivel 0,01 (bilateral).

Vamos a realizar el ajuste por mínimos cuadrados. Analizar > Regresión > Lineales > Estadísticos > Coeficientes de regresión: estimaciones; intervalos de confianza; Ajuste del modelo > Residuos: Durbin-Watson; Diagnósticos por caso; Valores atípicos fuera 3 desviaciones típicas





Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,886[a]	,786	,777	,08043	1,449

a. Variables predictoras: (Constante), Luz Principal

b. Variable dependiente: Canto Tablero

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,569	1	,569	88,021	,000[b]
	Residual	,155	24	,006		
	Total	,725	25			

a. Variable dependiente: Canto Tablero

b. Variables predictoras: (Constante), Luz Principal

Coeficientes[a]

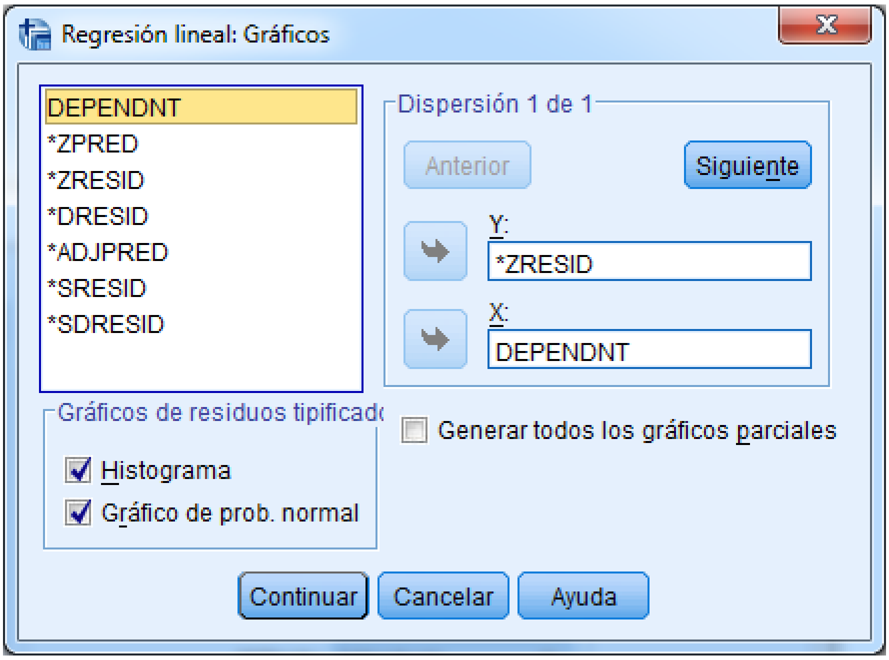
Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	,141	,078		1,814	,082	-,019	,301
	Luz Principal	,033	,004	,886	9,382	,000	,026	,040

a. Variable dependiente: Canto Tablero

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	,6421	1,1303	,8538	,15091	26
Residual	-,09283	,25109	,00000	,07880	26
Valor pronosticado tip.	-1,403	1,832	,000	1,000	26
Residuo típ.	-1,154	3,122	,000	,980	26

a. Variable dependiente: Canto Tablero


Regresión lineal: Gráficos
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Dispersión 1 de 1
Anterior
Siguiente
Y:
*ZRESID
X:
DEPENDNT
Gráficos de residuos tipificad
Histograma
Gráfico de prob. normal
Generar todos los gráficos parciales
Continuar
Cancelar
Ayuda

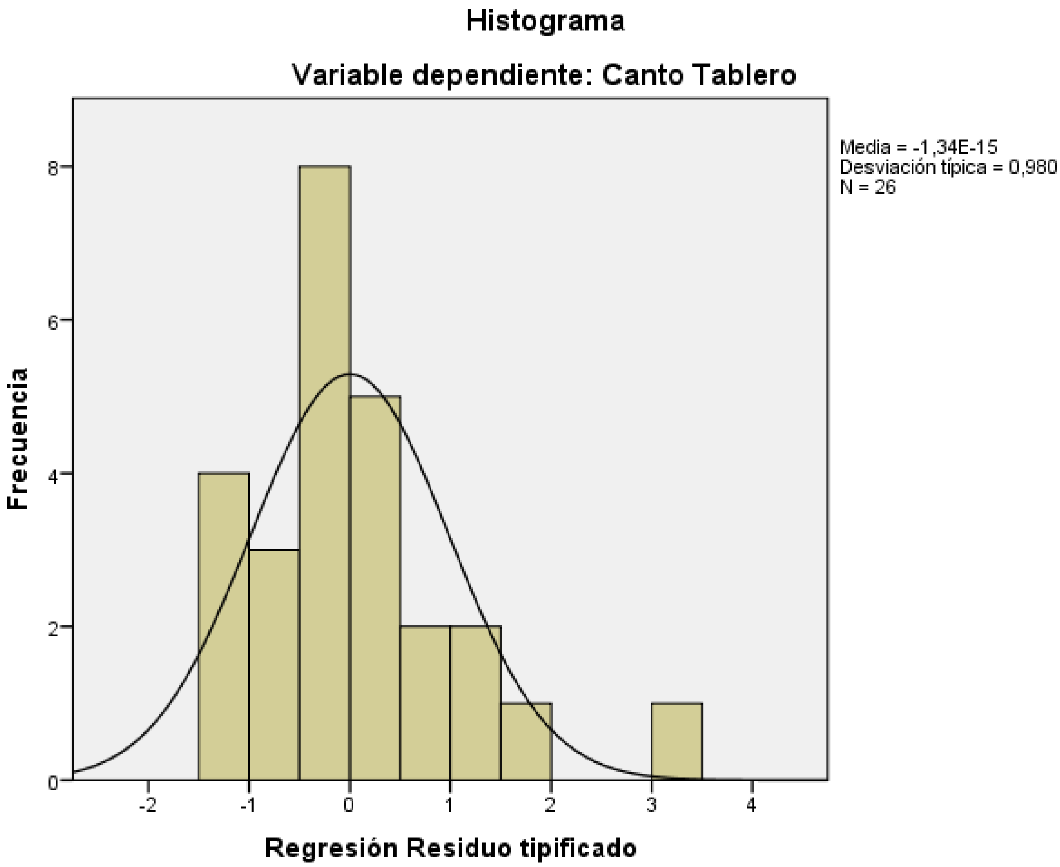

Histograma
Variable dependiente: Canto Tablero
Media = -1,34E-15
Desviación típica = 0,980
N = 26
Frecuencia
0
2
4
6
8
-2
-1
0
1
2
3
4
Regresión Residuo tipificado

Gráfico P-P normal de regresión Residuo tipificado

Variable dependiente: Canto Tablero

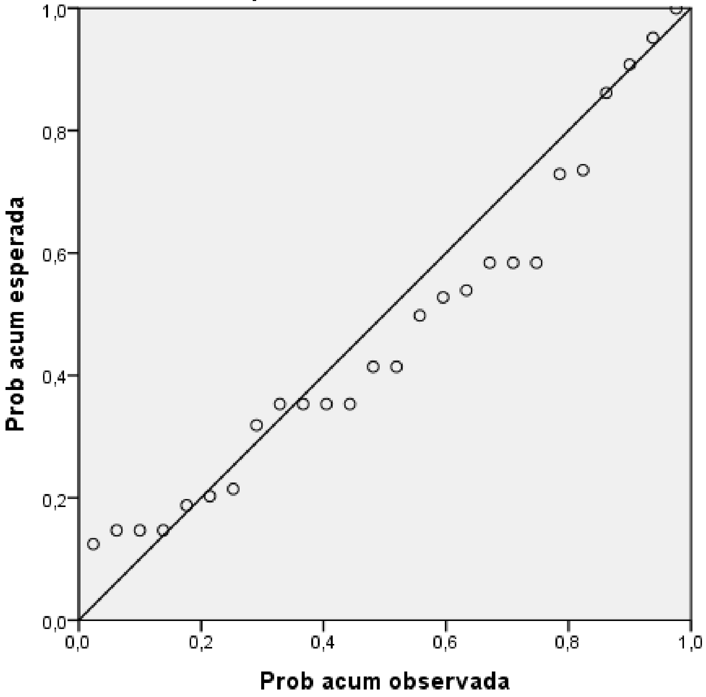
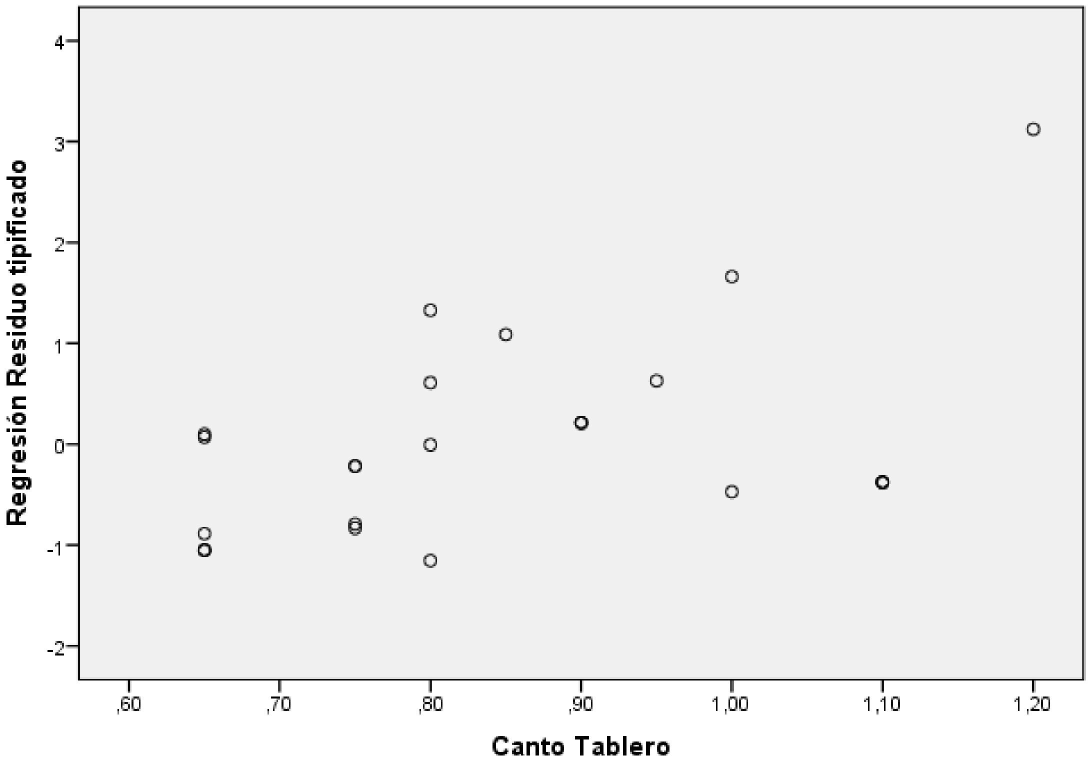
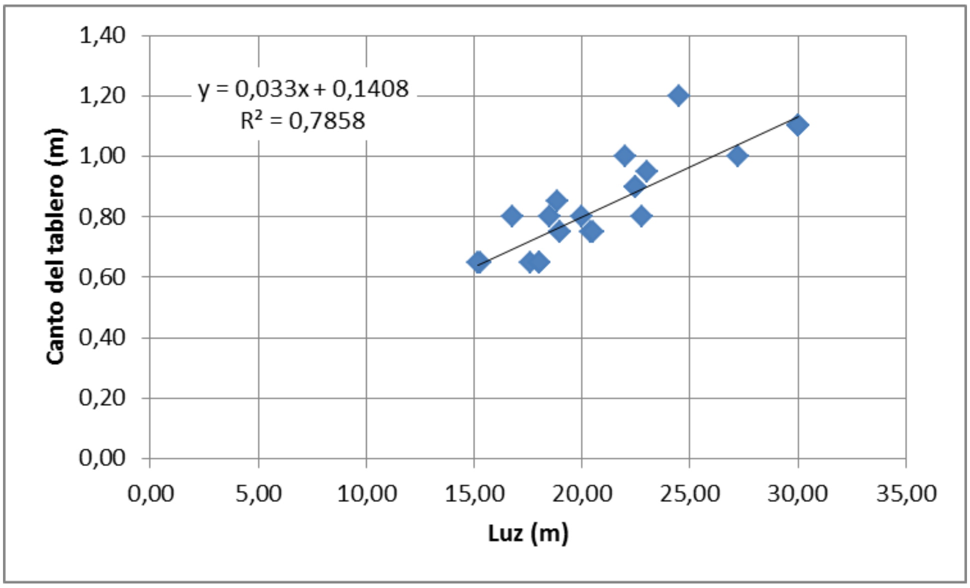


Gráfico de dispersión

Variable dependiente: Canto Tablero



Se puede realizar el análisis de regresión también con Excel. Se puede representar la gráfica y el ajuste. Insertar > Gráfico > Dispersión



Se puede comprobar que el R^2 que aparece no está ajustado.

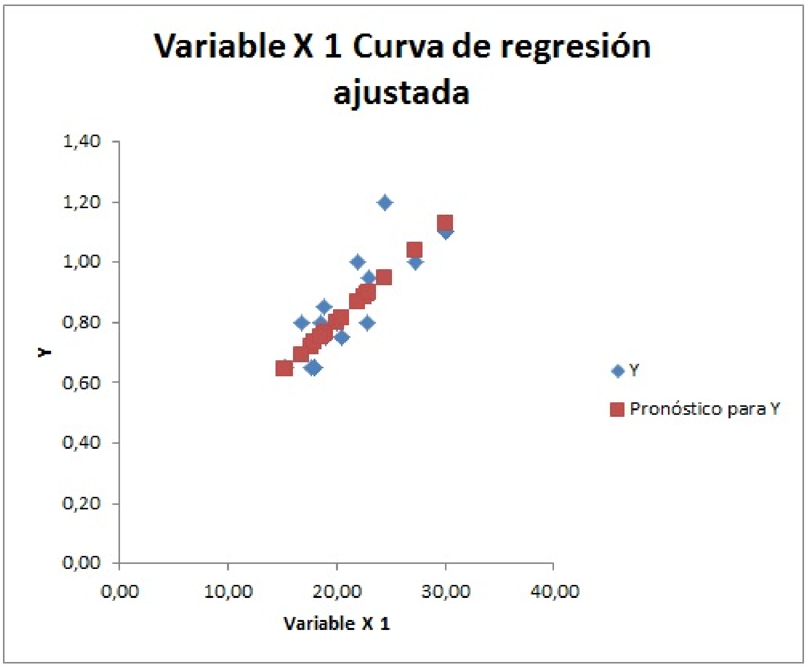
Es posible realizar un análisis más profundo con Excel. Para ello debemos tener instalado el complemento "Análisis de datos". Datos > Análisis de datos > Regresión > Nivel de confianza: 95%

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,88642845
Coeficiente de determinación R^2	0,7857554
R^2 ajustado	0,77682854
Error típico	0,08042723
Observaciones	26

ANÁLISIS DE VARIANZA

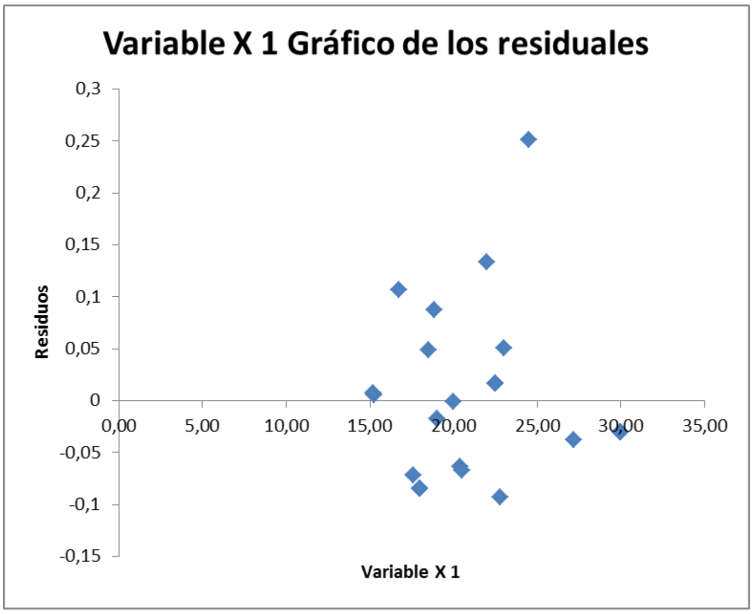
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	0,569370451	0,569370451	88,021493	1,6832E-09
Residuos	24	0,155244933	0,006468539		
Total	25	0,724615385			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	0,14078199	0,077623059	1,81366194	0,0822542	-0,019424132	0,30098811	-0,019424132	0,30098811
Variable X 1	0,03298463	0,003515744	9,38197703	1,683E-09	0,025728487	0,04024076	0,025728487	0,040240763



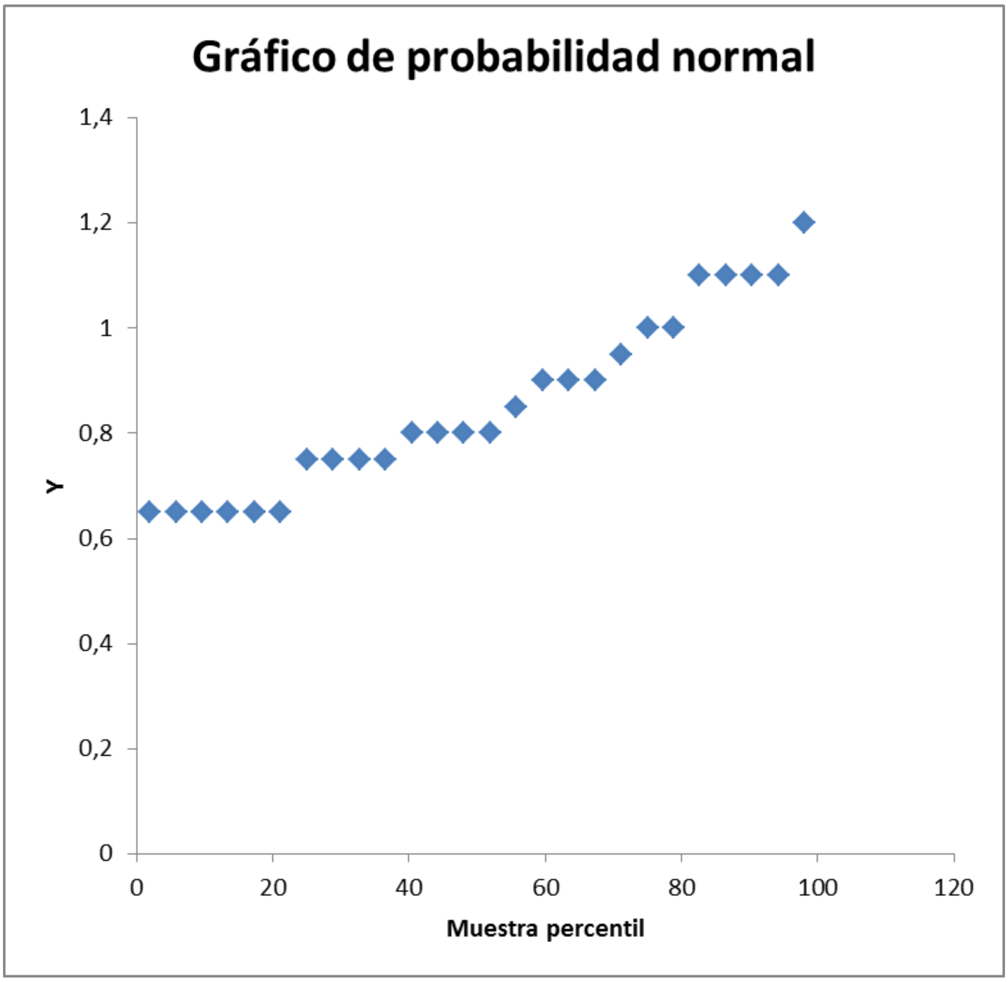
Análisis de los residuales

Observación	*Pronóstico para Y*	*Residuos*	*Residuos estándares*
1	1,130320745	-0,030320745	-0,384769958
2	1,130320745	-0,030320745	-0,384769958
3	1,130320745	-0,030320745	-0,384769958
4	1,130320745	-0,030320745	-0,384769958
5	0,882936056	0,017063944	0,216541282
6	0,767489868	-0,017489868	-0,221946251
7	0,882936056	0,017063944	0,216541282
8	0,734505243	-0,084505243	-1,0723707
9	0,734505243	-0,084505243	-1,0723707
10	0,721311392	-0,071311392	-0,904940872
11	0,800474493	-0,000474493	-0,006021311
12	0,948905306	0,251094694	3,186389201
13	0,767489868	-0,017489868	-0,221946251
14	0,644457216	0,005542784	0,070337878
15	0,642148292	0,007851708	0,099638098
16	0,734505243	-0,084505243	-1,0723707
17	0,813668343	-0,063668343	-0,807950649
18	1,037963794	-0,037963794	-0,48176018
19	0,816966806	-0,066966806	-0,849808107
20	0,892831444	-0,092831444	-1,178030109
21	0,762542174	0,087457826	1,109838955
22	0,693274461	0,106725539	1,354346043
23	0,750997555	0,049002445	0,621840545
24	0,882936056	0,017063944	0,216541282
25	0,866443743	0,133556257	1,694827587
26	0,899428369	0,050571631	0,641753507



Resultados de datos de probabilidad

Percentil	*Y*
1,923076923	0,65
5,769230769	0,65
9,615384615	0,65
13,46153846	0,65
17,30769231	0,65
21,15384615	0,65
25	0,75
28,84615385	0,75
32,69230769	0,75
36,53846154	0,75
40,38461538	0,8
44,23076923	0,8
48,07692308	0,8
51,92307692	0,8
55,76923077	0,85
59,61538462	0,9
63,46153846	0,9
67,30769231	0,9
71,15384615	0,95
75	1
78,84615385	1
82,69230769	1,1
86,53846154	1,1
90,38461538	1,1
94,23076923	1,1
98,07692308	1,2



Vamos a realizar el análisis con Matlab. Asignamos los datos de la luz a un vector columna X con "unos" en la primera columna y los del canto a un vector columna Y. Supondremos un nivel de confianza del 95%.

X =

```
    1.0000   30.0000

    1.0000   30.0000

    1.0000   30.0000

    1.0000   30.0000

    1.0000   22.5000

    1.0000   19.0000

    1.0000   22.5000

    1.0000   18.0000

    1.0000   18.0000
```

1.0000	17.6000
1.0000	20.0000
1.0000	24.5000
1.0000	19.0000
1.0000	15.2700
1.0000	15.2000
1.0000	18.0000
1.0000	20.4000
1.0000	27.2000
1.0000	20.5000
1.0000	22.8000
1.0000	18.8500
1.0000	16.7500
1.0000	18.5000
1.0000	22.5000
1.0000	22.0000
1.0000	23.0000

Y =

1.1000

1.1000

1.1000

1.1000

0.9000

0.7500

0.9000

0.6500

0.6500

0.6500

0.8000

1.2000

0.7500

0.6500

0.6500

0.6500

0.7500

1.0000

0.7500

0.8000

0.8500

0.8000

0.8000

0.9000

1.0000

0.9500

```
>> [b,bint,e,eint,stats]=regress(Y,X, 0.05)
```

Los coeficientes del modelo de mínimos cuadrados:

b =

0.1408

0.0330

Los intervalos de confianza para los coeficientes son los siguientes:

bint =

-0.0194	0.3010
0.0257	0.0402

El vector de residuales es:

e =

-0.0303

-0.0303

-0.0303

-0.0303

0.0171

-0.0175

0.0171

-0.0845

-0.0845

-0.0713

-0.0005

0.2511

-0.0175

0.0055

0.0079

-0.0845

-0.0637

-0.0380

-0.0670

-0.0928

0.0875

0.1067

0.0490

0.0171

0.1336

0.0506

El intervalo de confianza para los residuos:

eint =

-0.1840	0.1234
-0.1840	0.1234
-0.1840	0.1234
-0.1840	0.1234

-0.1489	0.1830
-0.1825	0.1475
-0.1489	0.1830
-0.2445	0.0755
-0.2445	0.0755
-0.2320	0.0894
-0.1663	0.1654
0.1265	0.3756
-0.1825	0.1475
-0.1539	0.1650
-0.1514	0.1671
-0.2445	0.0755
-0.2274	0.1001
-0.1982	0.1222
-0.2305	0.0966
-0.2540	0.0683
-0.0732	0.2481
-0.0489	0.2624
-0.1143	0.2123
-0.1489	0.1830
-0.0224	0.2896
-0.1140	0.2151

El coeficiente de determinación R^2, el valor del estadístico F, el valor p de la prueba F y la estimación del error de la varianza.

stats =

0.7858 88.0215 0.0000 0.0065

El coeficiente de correlación lineal sería:

```
>> corrcoef(X(:,2),Y)
```

ans =

```
    1.0000    0.8864

    0.8864    1.0000
```

Podemos dibujar también el gráfico de los puntos muestrales y la recta de regresión.

```
>> clf

>> scatter(X(:,2),Y,'filled'),grid on

>> hold on, ezplot('0.1408+0.0330*x',[10,35])

>> legend('Recta de regresión', 'Datos muestrales', 2)
```