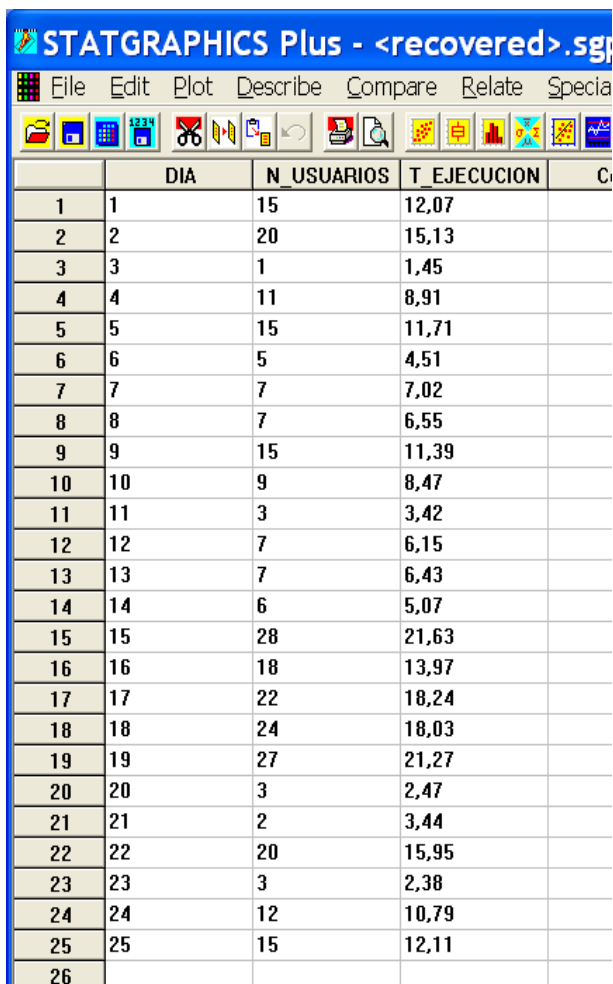


## PRÁCTICA 9. INTRODUCCIÓN A LA REGRESIÓN LINEAL

### Objetivo

En esta sesión de laboratorio se aplican los conceptos y herramientas relacionados con la recta de regresión. Se utiliza esta herramienta con *Statgraphics*.

Los datos son de un experimento para estudiar el tiempo que tarda un sistema informático en red en ejecutar un conjunto de instrucciones. Dicho tiempo depende, básicamente, del número de usuarios conectados a él. Para estudiar esta relación, se anotó durante 25 días el número de usuarios a las 9 de la mañana, y el tiempo que tardó en ejecutarse un programa prueba (*benchmark*). Los datos se recogen en el fichero **Tejecucion.sf3** disponible en **PoliformaT** (Figura 1).

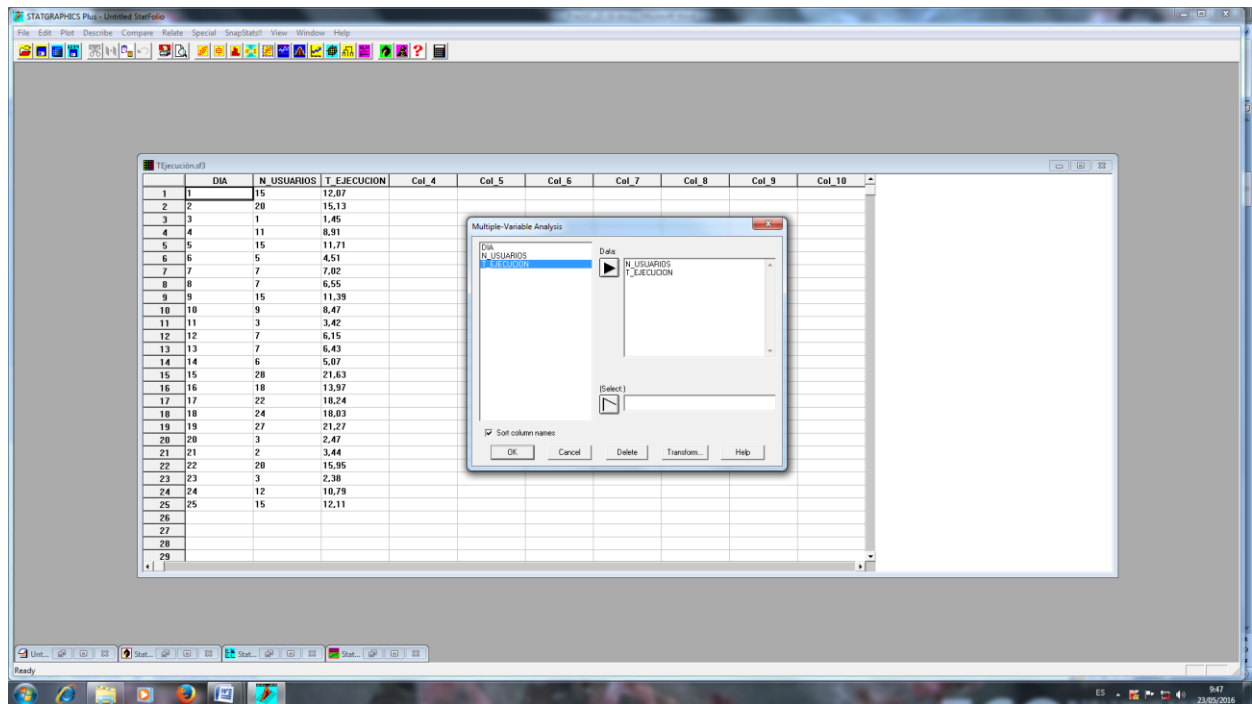


	DIA	N_USUARIOS	T_EJECUCION	C
1	1	15	12,07	
2	2	20	15,13	
3	3	1	1,45	
4	4	11	8,91	
5	5	15	11,71	
6	6	5	4,51	
7	7	7	7,02	
8	8	7	6,55	
9	9	15	11,39	
10	10	9	8,47	
11	11	3	3,42	
12	12	7	6,15	
13	13	7	6,43	
14	14	6	5,07	
15	15	28	21,63	
16	16	18	13,97	
17	17	22	18,24	
18	18	24	18,03	
19	19	27	21,27	
20	20	3	2,47	
21	21	2	3,44	
22	22	20	15,95	
23	23	3	2,38	
24	24	12	10,79	
25	25	15	12,11	
26				

Figura 1. Introducción de datos.

**Pregunta 1.** Representa el diagrama de dispersión entre **T\_EJECUCION** Y **N\_USUARIOS**, y calcula el coeficiente de correlación lineal  $r$  (Figura 2). ¿Cómo es la relación entre estas dos variables?

**SOLUCIÓN:** Hay una relación lineal positiva fuerte  $r=0,9955$

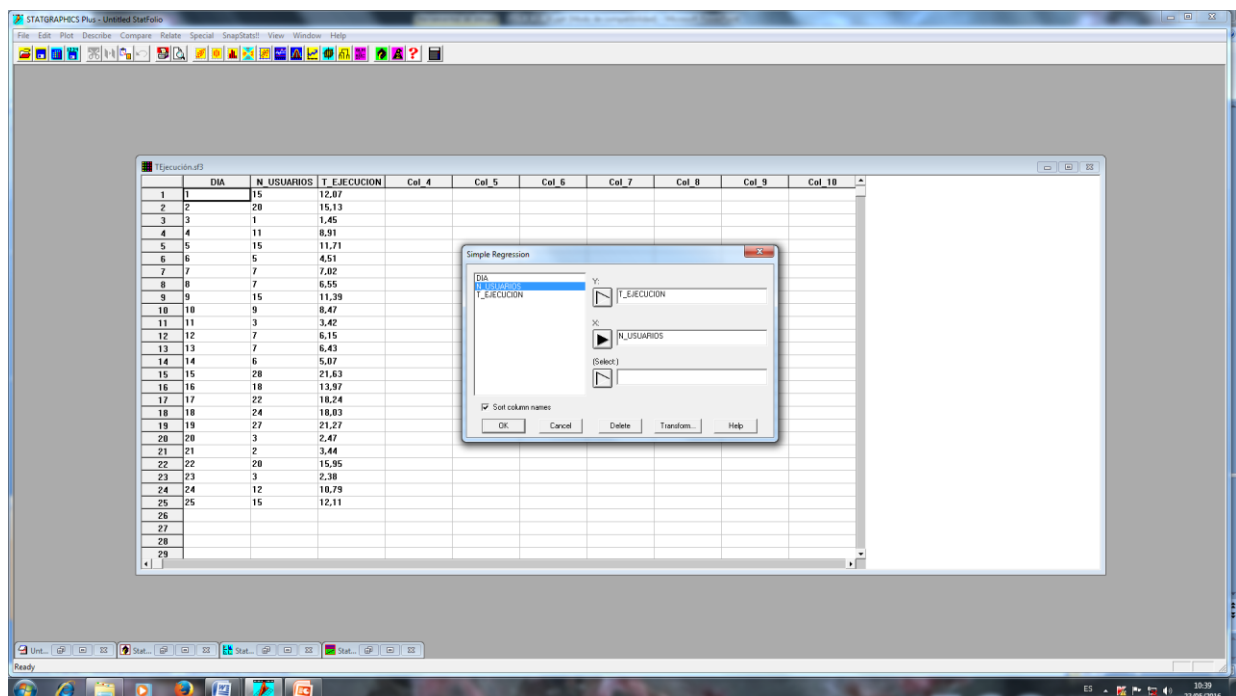


**Figura 2.** Describe....Numeric Data...Multiple Variable Analysis...

Data: N\_USUARIOS

T\_EJECUCIÓN

**Pregunta 2.** Plantea el modelo de regresión. Calcula los parámetros de la recta de regresión (ordenada y pendiente), y estudia su significación ( $\alpha=5\%$ ) (**Figuras 3 y 4**).



**Figura 3.** Relate....Simple Regression

Y:T\_EJECUCION

X:N\_USUARIOS

Regression Analysis - Linear model:  $Y = a + b \cdot X$ 

Dependent variable: T\_EJECUCION

Independent variable: N\_USUARIOS

Tipo de relación

Parameter		Estimate	Standard Error	T Statistic	P-Value
Intercept	<b>a</b>	1,04573	<b>S<sub>a</sub></b> 0,21043	4,96951	0,0001 [ <b>&lt; 0,05</b> ]
Slope	<b>b</b>	0,736479	<b>S<sub>b</sub></b> 0,014546	50,6311	0,0000 [ <b>&lt; 0,05</b> ]

Obtención de a y b

Figura 4. Estimación de los parámetros con Statgraphics.

**SOLUCIÓN:**

Modelo:

$$E(T\_EJECUCION/N\_USUARIOS) = \alpha + \beta N\_USUARIOS$$

Estimación:

$$E(T\_EJECUCION/N\_USUARIOS) = 1,04573 + 0,736479 \times N\_USUARIOS$$

Como el **p-value es menor que 0,05** ( $\alpha$ ) para  $H_0: \beta=0$ , el efecto de la variable explicativa (N\_USUARIOS) sobre el tiempo medio de ejecución es significativo. La ordenada del modelo **a** difiere significativamente de cero (p-value <0,05).

El **ANOVA** del modelo estimado es:

Analysis of Variance						
Source		Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	(SCE)	859,077	1	859,077	2563,51	0,0000 [ <b>&lt; 0,01</b> ]
Residual	(SCR)	7,70771	23	(CMR) 0,335118		
Total (Corr.)	(SCT)	866,785	24			

Correlation Coefficient = 0,995544 → Coeficiente de correlación (r)

R-squared = 99,1108 percent → Coeficiente de Determinación ( $R^2$ ) = (SCE/SCT) × 100

R-squared (adjusted for d.f.) = 99,0721 percent

Standard Error of Est. = 0,578894 → Desviación Típica Residual ( $S_R$ ) =  $\sqrt{\text{CMR}}$ Figura 5. ANOVA del modelo y coeficiente de determinación  $R^2$

**Pregunta 3.** ¿Cuál es la conclusión del ANOVA ( $\alpha=1\%$ )? ¿Cuánto vale el coeficiente de determinación  $R^2$  y la desviación típica residual?

**SOLUCIÓN:**

La hipótesis  $\beta=0$  se rechaza porque  $p\text{-value}$  es  $< 0,01$ . El coeficiente de determinación  $R^2$ , en este caso es 99,1108%: porcentaje de la variación lineal del tiempo de ejecución (variable dependiente), asociada a los cambios en el número de usuarios que trabajan en el sistema (variable independiente).

El orden de magnitud del efecto que conjuntamente tienen sobre la variable dependiente otras variables que pueden influir, en mayor o menor medida, sobre la v.a. dependiente y que no se han tenido en cuenta en la recta de regresión, está recogido en la **varianza residual**.

En este ejemplo,  $S^2_{\text{residual}} = 0,3351$  segundos<sup>2</sup> (CMresidual)

Y  $S_{\text{residual}} = 0,5789$  segundos.

**Pregunta 4.** ¿Qué interpretación práctica tienen los valores **a** y **b**?

**SOLUCIÓN:**

$b=0,7364$  es la pendiente de la recta de regresión. Expresa en cuánto aumenta ( $b>0$ ) el tiempo medio de ejecución cuando el número de usuarios conectados se incrementa en uno.

$a=1,0466$  Representa el tiempo medio de ejecución cuando no hay ningún usuario conectado.

**Pregunta 5.** ¿Entre que valores esta con una probabilidad del 55% el tiempo de ejecución en aquellos días en los que hay 10 usuarios?

**SOLUCIÓN:**

EL  $T_{\text{EJECUCION}} / (N_{\text{USUARIOS}} = 10)$  sigue una distribución normal con media:

$$E(T_{\text{EJECUCION}} / N_{\text{USUARIOS}} = 10) = 1,04573 + 0,736479 \times 10 = 8,41052 \text{ s}$$

y varianza=CMresidual=0,3351

desviación típica residual= 0,5789

El 55% de los valores del tiempo de ejecución cuando el número de usuarios es 10 estarán en el intervalo  $[8,41-Z \cdot 0,5789, 8,41+Z \cdot 0,5789]$

$Z$  es el valor de la normal tipificada que,  $P(N(0,1)>Z)=(1-0,55)/2$

$$P(N(0,1)>Z)=0,225 \Rightarrow Z \approx 0,76$$

(con Statgraphics  $Z=0,7554$ )

El intervalo es:  $[8,41-0,76 \cdot 0,5789, 8,41+0,76 \cdot 0,5789]=[7,97, 8,85]$  segundos

**Pregunta 6.** Calcula la media, la varianza y los coeficientes de asimetría y curtosis estandarizados de los residuos (**Figuras 6 y 7**). Represéntalos en Papel Probabilístico Normal (**Figura 8**). ¿Se distribuye la variable tiempo de ejecución normalmente?

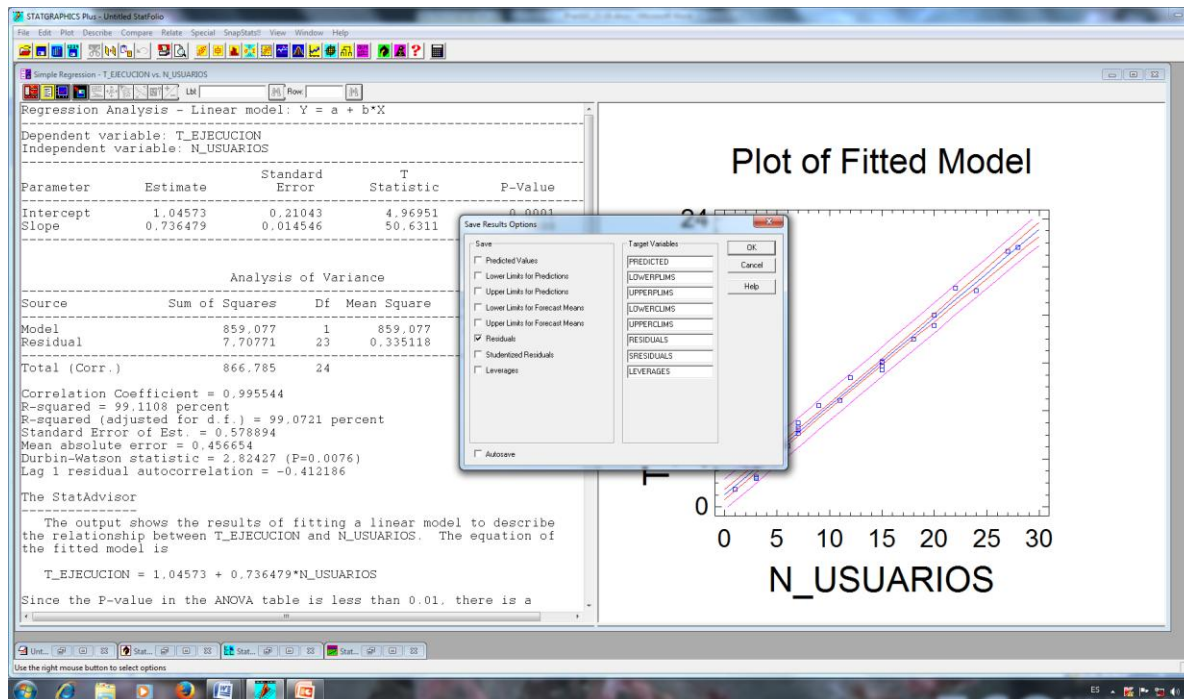


Figura 6. Cuadro de diálogo para calcular los residuos con Statgraphics.

**Summary Statistics for Residuos**

Count = 25  
 Average =  $-7,2E-8$   
 Median =  $-0,0371487$   
 Mode =  
 Variance =  $0,321154$   
 Standard deviation =  $0,566705$   
 Minimum =  $-0,875169$   
 Maximum =  $0,991726$   
 Range =  $1,8669$   
 Lower quartile =  $-0,382919$   
 Upper quartile =  $0,33933$   
 Std. skewness =  $0,68251$   
 Std. kurtosis =  $-0,871726$

Figura 7. "Summary Statistics" de los Residuos

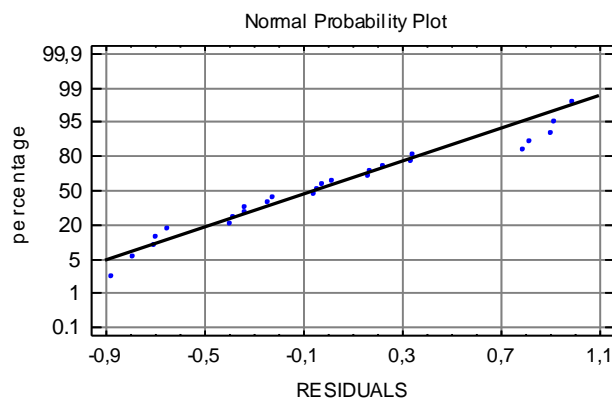


Figura 8. Residuos sobre PPN

La media es  $-7,2 \cdot 10^{-8}$  ( $\approx 0$ ), la varianza ( $S_u^2$ ) es  $0,321154$ . Los valores de asimetría y curtosis (estándar) son  $0,68251$  y  $-0,871726 \in [-2, 2]$ . La representación de los residuos en PPN indica que siguen distribución normal, por lo que el tiempo de ejecución se puede estudiar también con este modelo.