

PRÁCTICAS 9 - 10. ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL II INTRODUCCIÓN A LOS MODELOS DE REGRESIÓN LINEAL

Objetivo

El objeto de esta sesión de laboratorio es repasar y reforzar los conceptos y herramientas vistos en clase sobre las variables aleatorias bidimensionales (2ª Parte de la UD 2) y la recta de regresión (4ª parte de la UD5). En concreto el diagrama de dispersión, los parámetros de la covarianza y coeficiente de correlación lineal y los aspectos más elementales e importantes de los modelos de regresión. Además, se pretende que el alumno aprenda a utilizar estas herramientas mediante software estadístico adecuado como el *Statgraphics*.

Los datos a utilizar son los relativos a un experimento efectuado con el fin de estudiar el rendimiento de un sistema informático (SI) en red.

NOTA: Se recomienda que, durante el trabajo no presencial del alumno, los resultados obtenidos a partir del Statgraphics en esta práctica se calculen "a mano" y se cotejen con los mismos.

1. Introducción de los datos.

Se sabe que el tiempo (segundos) que tarda un sistema informático en red en ejecutar un conjunto de instrucciones depende, básicamente, del número de usuarios conectados a él.

Para estudiar la posible relación entre ambas características se anotó durante 25 días el número de usuarios a las 9 de la mañana y el tiempo en que el sistema tardó en ejecutar cierto programa prueba (*benchmark*). Los datos constatados se recogen en el fichero **Pract9_EST-GII-ETSINF_17-18.sf3** disponible en **PoliformaT** (Figura 1).

2. Análisis descriptivo.

El primer paso, y no menos importante, de cualquier estudio estadístico es el análisis descriptivo o exploratorio de los datos.

Suponiendo que ya se ha llevado a cabo una exploración de cada una de las componentes de la v.a. bidimensional, comenzaremos por obtener el diagrama de dispersión entre **T_EJECUCION** Y **N_USUARIOS** (Figura 2).

Pregunta 1. A la vista del diagrama de dispersión, responde si es posible que exista algún tipo de relación entre ambas componentes de la v.a. bidimensional y en caso afirmativo, di de que tipo es (lineal, exponencial,...). ¿Cuál crees que será el valor aproximado del coeficiente de correlación lineal (r)?

Seguidamente, obtendremos la matriz de varianzas-covarianzas. Para ello seguimos la secuencia: **Describir > Datos Numéricos > Análisis Multivariado**

Introducimos el nombre de las dos variables y desde el menú de **Tablas y Gráficos**, seleccionamos la opción **Covarianzas**.

Pregunta 2. Utilizando exclusivamente la información contenida en la matriz de varianzas-covarianzas, calcula cuál será la desviación típica para T_EJECUCION y N_USUARIOS y el coeficiente de correlación entre ambas. Compara el valor de r obtenido con el adelantado en la pregunta anterior.

Finalmente, solicitamos la matriz de correlación.

Pregunta 3. A la vista de la matriz de correlaciones, compara el resultado avanzado en la pregunta anterior con el valor de r que da el *Statgraphics*.

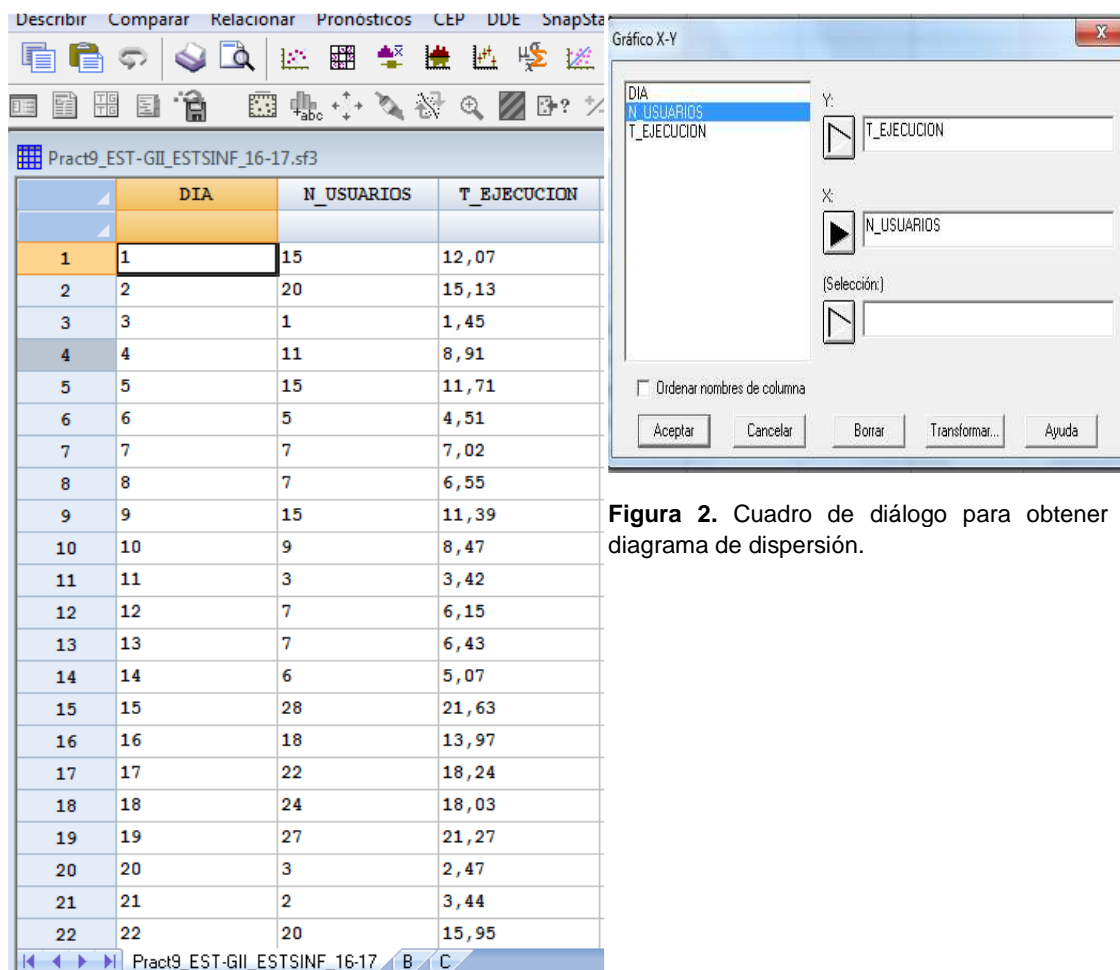


Figura 2. Cuadro de diálogo para obtener el diagrama de dispersión.

Figura 1. Introducción de datos.

Pregunta 4. ¿Tendría algún interés en este caso obtener la expresión de la recta de regresión que relaciona las variables? ¿Sería adecuado?

3. Obtención de la ecuación de la recta de regresión

NOTA. Ayúdate de la información del panel “Summary Statistics” y de los resultados obtenidos en las Preguntas 2 y 3.

Pregunta 5. Con los datos de los que hasta ahora dispones, calcula “A MANO” los coeficientes de la ecuación de la recta de regresión.

Para obtener los coeficientes **a** y **b** de la recta de regresión mediante el *Statgraphics*, seleccionamos la opción de menú **Relacionar > Un Factor > Regresión**, (Figura 3) y a continuación indicamos cuál es la variable dependiente (Y) y la independiente (X) (Figura 4). La salida correspondiente y el significado de la información más relevante se muestran en la Figura 5.

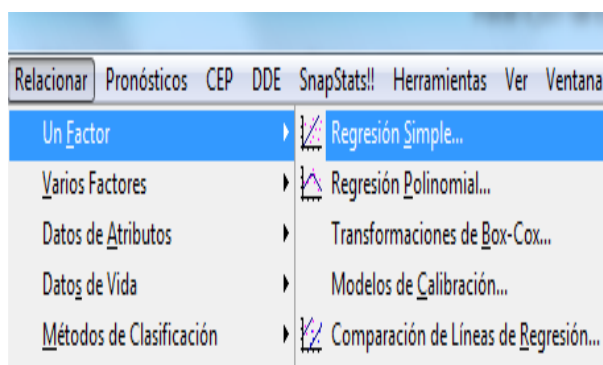


Figura 3. Opción de menú para realizar un análisis de regresión simple.

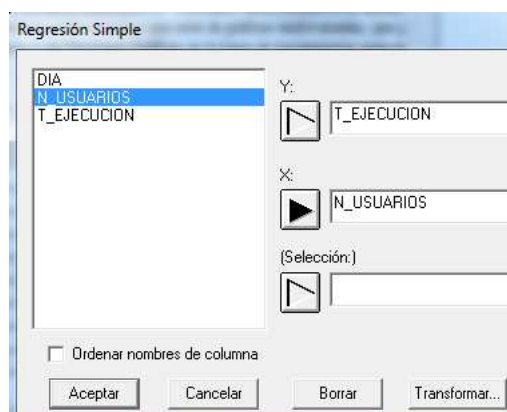


Figura 4. Selección de las variables dependiente e independiente del modelo.

Pregunta 6. ¿Cuál es la recta de regresión que relaciona el número de usuarios y el tiempo de ejecución en este ejemplo?

Pregunta 7. ¿Qué tiempo de ejecución cabe esperar en promedio para aquellos días en los que hay 15 usuarios trabajando en el SI?

Pregunta 8. ¿Qué interpretación práctica tienen los valores **a** y **b**?

Pregunta 9. Según el modelo obtenido (recta de regresión), si a las 9 de la mañana el número de usuarios aumenta de 15 a 16, entonces ¿cómo se modifica el tiempo de ejecución esperado?

Pregunta 10. ¿Qué tiempo de ejecución se ha constatado efectivamente los días 1, 5, 9 y 25 en los que el número de usuarios ha sido de 15? (ver el editor de datos) ¿Hay diferencia entre estos valores y el obtenido en la **Pregunta 7**? Si existen diferencias calcúlalas y explica que representan.

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

Variable Dependiente: T_EJECUCION

Variable Independiente: N_USUARIOS

Tipo de relación

Parametro	Estimación	Error Standard	T Estadístico	P-Valor
Intercept	a 1,04573	0,21043	4,96951	0,0001 [< 0,05]
Slope	b 0,736479	0,014546	50,6311	0,0000 [< 0,05]

Figura 5. Salida del Statgraphics. Estimación de los parámetros.

Obtención de **a** y **b**

Según los datos de la tabla (**Figura 5**) los **errores de la estimación** de **a** y **b** son 0,21 (**S_a**) y 0,014 (**S_b**), respectivamente.

Como el **p-value** es menor que **0,05** (α) en ambos casos, podemos decir que el efecto de la variable explicativa (N_USUARIOS) es significativo y que **a** y **b** sí **aparecen en la ecuación final del modelo**.

4. Significación global del ajuste. Calidad del ajuste

Para determinar la calidad del ajuste, esto es, determinar lo bien que el modelo (una recta en nuestro caso) representa nuestra realidad observada, lo primero que hay que hacer es cerciorarnos de si las variables explicativas tienen efecto real a nivel poblacional. A esta primera pregunta se contesta mediante la realización de un **ANOVA** que el *Statgraphics* efectúa por defecto al mismo tiempo que estima los parámetros del modelo (**Figura 6**).

Análisis de Varianza

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	F-Ratio	P-Valor
Model	(SCE) 859,077	1	859,077	2563,51	0,0000 [< 0,05]
Residual	(SCR) 7,70771	23	(CMR) 0,335118		
Total (Corr.)	(SCT) 866,785	24			

Coefficiente Correlación = 0,995544 → Coeficiente de correlación (**r**)

R-cuadrado = 99,1108 por ciento → Coeficiente de Determinación (**R²**) = (**SCE/SCT**)x100

R-cuadrado (ajustado) = 99,0721 por ciento

Error Standard de Est. = 0,578894 → Desviación Típica Residual (**S_R**) = $\sqrt{\text{CMR}}$

Figura 6. Salida del Statgraphics (ANOVA)

Como el **p-value es menor que 0,05 (α)**, podemos decir que las variables explicativas del modelo, tomadas conjuntamente, tienen un efecto significativo sobre el tiempo medio de ejecución. Como en este caso sólo hay una variable independiente, esto quiere decir que el efecto que tiene el número de usuarios (N_USUARIOS) sobre el tiempo medio de ejecución (T_EJECUCIÓN) es significativo, el modelo (la recta) es globalmente significativo, pero, ¿cuál es la calidad de este ajuste?

RECORDAR. En el caso de **Regresión Lineal Simple**, el Coeficiente de Determinación (R^2) y la estimación de la Varianza Residual (S^2_R) se pueden calcular como:

$$R^2 = r^2 \times 100 \text{ y } S^2_R = S^2_Y (1 - r^2_{XY})$$

En cuanto a la **calidad del ajuste** (lo bueno o malo que es el modelo en sus predicciones) recordad que éste se cuantifica mediante el **Coeficiente de determinación R^2** . Así, volviendo a la ventana del ANOVA (**Figura 6**):


Pregunta 11. ¿Qué porcentaje de la variación del tiempo de ejecución (variable dependiente) está provocada o explicada por los cambios en el número de usuarios que trabajan en el sistema (variable independiente)?

Pregunta 12. ¿Cuál es el orden de magnitud del efecto conjunto de aquellos factores no considerados en el modelo?

RECORDAR: los valores observados para cada posible valor de la variable independiente (X_i), siguen, a nivel poblacional, una distribución condicional que se asume normal. La media de cada distribución condicional es el valor que se obtiene a partir de la recta de regresión para cada X_i ($E(Y/X=X_i) = a + b x_i$) y se acepta que la varianza es la misma para todas las distribuciones condicionales y coincide con la **varianza residual**.

Pregunta 13. ¿Entre que valores estará, en el 95% de los casos, el tiempo de ejecución en aquellos días en los que hay 10 usuarios utilizando el sistema informático a las 9 de la mañana?

5. Validación del modelo. Análisis de los residuos.

Para disponer de los residuos para cada valor de la variable dependiente (T_EJECUCION), pulsamos en “**Opciones para Guardar Resultados**”,  seleccionamos “**Residuos**” (**Figura 7**). Así, el *Statgraphics* crea una nueva variable aleatoria que por defecto le denomina **Residuos**, aunque se le puede cambiar el nombre introduciéndolo en el campo “**Variable Destino**” (**Figura 8**).

En este cuadro de diálogo también es posible obtener otros subcálculos del *Statgraphics*. Seleccionando “**Predicciones**” también se pueden guardar los valores que se predicen para cada valor de la variable independiente, N_USUARIOS, según el modelo descrito por la recta de regresión.

Pregunta 14. ¿Cuáles son los residuos para los días 1, 5, 9 y 25? ¿Qué relación guardan estos valores con los obtenidos en la **Pregunta 10**?

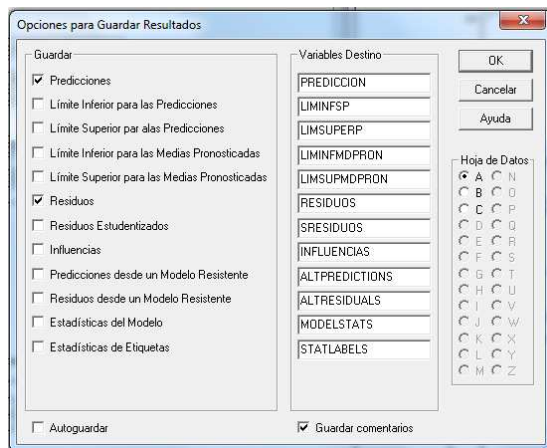


Figura 7. Cuadro de diálogo para seleccionar los residuos calculados por el *Statgraphics*.

DIA	N_USUARIOS	T_EJECUCION	PREDICCION	RESIDUOS
1	15	12,07	12,0929	-0,0229192
2	20	15,13	15,7753	-0,645315
3	1	1,45	1,78221	-0,332211
4	11	8,91	9,147	-0,237002
5	15	11,71	12,0929	-0,382919
6	5	4,51	4,72813	-0,218127
7	7	7,02	6,20109	0,818914
8	7	6,55	6,20109	0,348914
9	15	11,39	12,0929	-0,702919
10	9	8,47	7,67404	0,795956
11	3	3,42	3,25517	0,164831
12	7	6,15	6,20109	-0,0510857
13	7	6,43	6,20109	0,228914
14	6	5,07	5,46461	-0,394607
15	28	21,63	21,6671	-0,0371487
16	18	13,97	14,3024	-0,332357
17	22	18,24	17,2483	0,991726
18	24	18,03	18,7212	-0,691232
19	27	21,27	20,9307	0,33933
20	3	2,47	3,25517	-0,785169
21	2	3,44	2,51869	0,92131
22	20	15,95	15,7753	0,174685
23	3	2,38	3,25517	-0,875169
24	12	10,79	9,88348	0,906518
25	15	12,11	12,0929	0,0170808

Figura 8. Ventana del conjunto de datos en la que se muestra la nueva variable **Residuos**.

En un estudio de regresión, la realización de un **análisis** exploratorio o **descriptivo de los residuos** es de gran importancia, permite, entre otras cosas, detectar posibles anomalías y pautas de comportamiento no lineales entre las componentes de la v.a. bidimensional.

Comenzaremos calculando los principales parámetros de posición, dispersión. etc.

NOTA. Recordar que hay que elegir la opción de menú **Describir > Datos Numéricos > Análisis de Una Variable...** y seleccionar la variable **Residuos** para el análisis.

Pregunta 15. Obtener la media, la varianza y los coeficientes de asimetría y curtosis de los residuos (**Figura 9**), representarlos en Papel Probabilístico Normal y determinar si podríamos considerar que estos datos proceden de una población que sigue una distribución normal.

Resumen Estadístico para B.RESIDUOS

Recuento	25
Promedio	-7,2E-8
Mediana	-0,0371487
Varianza	0,321154
Desviación Estándar	0,566705
Mínimo	-0,875169
Máximo	0,991726
Rango	1,8669
Sesgo Estandarizado	0,68251
Curtosis Estandarizada	-0,871726

Figura 9. Resumen de los parámetros descriptivos para la variable Residuos.

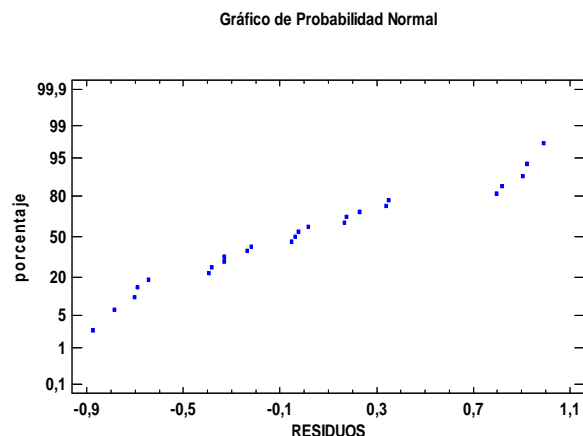


Figura 10. Residuos sobre PPN.

Respuestas a las preguntas propuestas

Pregunta 1

Como era previsible a partir del conocimiento previo sobre las características a estudiar, se observa claramente que los puntos se alinean en forma de recta, por lo que es altamente probable que exista una relación lineal positiva o directa entre ambas variables, esto es, a mayor número de usuarios ejecutando el *benchmark*, mayores valores se obtienen del tiempo de ejecución.

Así, podríamos avanzar que el valor aproximado de r estará entre 0,8 y 1, muy posiblemente alrededor de 0,9.

Pregunta 2

Las desviaciones típicas son 8,1236 usuarios para $N_USUARIOS$ y 6,0096 segundos para $T_EJECUCION$.

$$S_x = \sqrt{S_x^2} = \sqrt{65,9933} = 8,1236 \text{ usuarios} \quad S_y = \sqrt{S_y^2} = \sqrt{36,116} = 6,0096 \text{ segundos}$$

$$r_{xy} = \frac{\text{cov}_{xy}}{S_x S_y} = \frac{48,6027}{8,1236 \times 6,0096} = 0,9955$$

Como se había avanzado en la pregunta anterior, r está alrededor de 0,9

Pregunta 3

El valor de r que da el Statgraphics es 0,9955, tal y como se ha calculado en la pregunta anterior.

Pregunta 4

Como confirma el valor de r , existe una fuerte relación lineal positiva entre las dos componentes de la v.a. bidimensional, con lo que, en este caso, sí resultaría adecuado obtener la expresión matemática que modela dicha relación (recta de regresión). De este modo podríamos predecir los tiempos de ejecución, en promedio, en función del número de usuarios que estén utilizando el SI en red.

Pregunta 5

$$b = r_{xy} \frac{S_y}{S_x} = 0,9955 \frac{6,0096}{8,1236} = 0,7364 \quad a = \bar{Y} - b\bar{X} = 9,9424 - 0,7364 \times 12,08 = 1,0466$$

Pregunta 6

$$E(T_EJECUCION/N_USUARIOS) = 1,04573 + 0,736479 \times N_USUARIOS$$

Pregunta 7

$$E(T_EJECUCION/N_USUARIOS=15) = 1,04573 + 0,736479 \times 15 = 12,0929 \text{ s (t)}$$

Pregunta 8

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

$a=1,0466$ es el punto por donde la recta de regresión corta el eje Y. Representa el tiempo de ejecución cuando no hay ningún usuario conectado. En este ejemplo es posible que aún cuando no haya nadie trabajando hay ciertas instrucciones que se ejecutan en el sistema igualmente.

Pregunta 9

De acuerdo con la interpretación de **b** dada en la pregunta anterior, el aumento sería de 0,7364 s

Pregunta 10

$$t_1 = 12,07 \text{ s.} \quad t_5 = 11,71 \text{ s.} \quad t_9 = 11,39 \text{ s.} \quad t_{25} = 12,11 \text{ s.}$$

El valor que ha estimado la recta (**Pregunta 7**) ha sido 12,0929 y es diferente, aunque muy parecido, a los valores t_1 , t_5 , t_9 y t_{25}

La discrepancia entre los valores representa las diferencias entre lo observado y lo estimado, que son:

$$t_1 - t = -0,0229 \text{ s; } t_5 - t = -0,3929 \text{ s; } t_9 - t = -0,7029 \text{ s; } t_{25} - t = 0,01708$$

Pregunta 11

Esto es lo que representa el **coeficiente de determinación R^2** , que en este caso es el 99,1108% (**Figura 6**), lo que implica un muy buen ajuste.

Pregunta 12

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, bien porque no se ha considerado oportuno, bien porque su efecto no es controlable, y por tanto no cuantificable, bien como consecuencia de alguna anomalía, está recogido en la **varianza residual**.

En este ejemplo, $S^2_{\text{residual}} = 36,116 (1-(0,9955)^2) = 0,3243 \text{ s}^2$ (Comparar este valor con el que aparece en la **Figura 9** (*Variance*))

Y $S_{\text{residual}} = 0,5694 \text{ s.}$ (Comparar este valor con el que aparece en la **Figura 6** (*Standard Error of Est.*))

Pregunta 13

En promedio, $T_EJECUCION / (N_USUARIOS = 10)$ es el valor medio de una distribución condicional que sigue una distribución normal y cuya varianza es la varianza residual.

Por lo tanto, aproximadamente el 95% de los valores del tiempo de ejecución para aquellos días en los que el número de usuarios ha sido de 10 estarán en el intervalo $[m-2\sigma, m+2\sigma]$.

Donde $m = E(T_EJECUCION / N_USUARIOS = 10) \rightarrow$ lo que se obtiene a partir de la recta y σ es la varianza residual (σ_R)

$$m = T_EJECUCION / (N_USUARIOS = 10) = 1,04573 + 0,736479 \times 10 = 8,41052 \text{ s}$$

S (estimación de la σ_R) = Desviación típica residual (*Standard Error of Est.*) = S_{residual} = 0,5694 s

$[m - 2S_R, m + 2S_R] \rightarrow [8,41052 - 2 \times 0,5694, 8,41052 + 2 \times 0,5694] \rightarrow$ Aproximadamente, en el 95% de los casos en los que hay 10 usuarios, el tiempo de ejecución estará entre 7,27172 y 9,54932 segundos.

Pregunta 14

$$u_1 = -0,0229192; \quad u_5 = -0,382919; \quad u_9 = -0,702919; \quad u_{25} = 0,0170808$$

Estas diferencias entre lo realmente observado y lo que el modelo predice son los **residuos**. Son precisamente los valores que se han calculado en la pregunta 10.

Pregunta 15

La media es prácticamente 0, la varianza (S^2_u) es 0,321154 y los valores para los parámetros de asimetría y curtosis (estándar) son 0,68251 y -0,871726 respectivamente (**Figura 9**)

Como *Std. skewness* y *Std. Kurtosis* toman valores en el intervalo $[-2, 2]$, nada impide pensar que los datos de los residuos proceden de una población normal (simétrica y de apuntamiento normal)

Al representar los residuos sobre PPN tampoco se observan indicios de que éstos no procedan de una distribución normal, ya que los puntos forman aproximadamente una recta (**Figura 10**).

Esta obra está bajo una licencia Reconocimiento-No comercial-Compartir bajo la misma licencia 2.5 España de Creative Commons. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-sa/2.5/es/>

