Regresión Lineal Múltiple con StatGraphics Centurion XVI

- **Objetivo**

  Estudiar relación entre una variable cuantitativa y un conjunto de variables predictoras cuantitativas.

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + \varepsilon$$

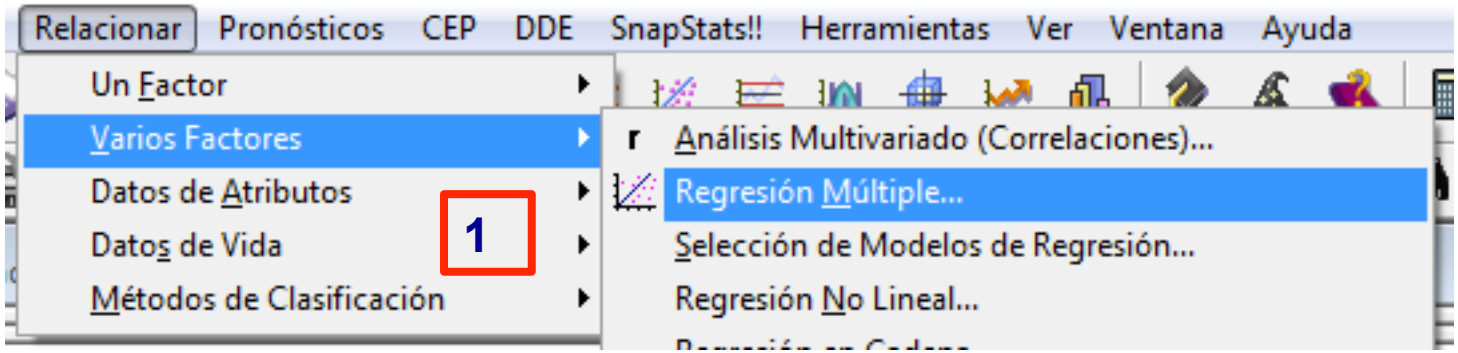
- **Hipótesis del modelo**
  1. **Normalidad:** los errores tienen una distribución Normal ($e_i \sim N(0; \sigma^2)$)
  2. **Linealidad:** Le esperanza de los errores es cero ($E(e_i)=0$)
  3. **Homocedasticidad:** La varianza de los errores son iguales
  4. **Independencia:** los errores son independientes

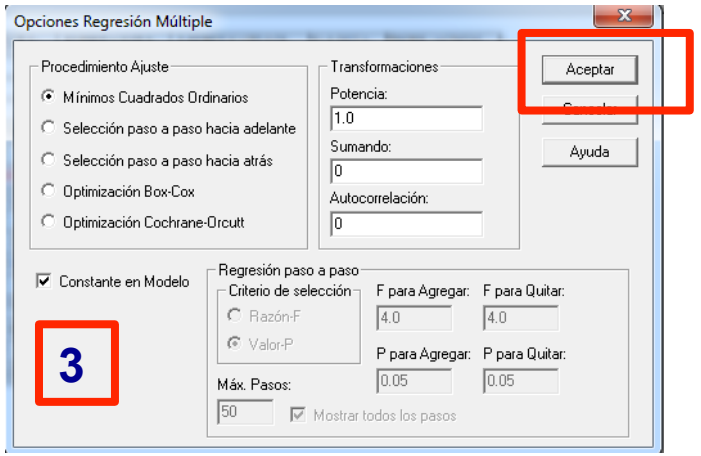
- **Requisitos adicionales de la regresión múltiple**

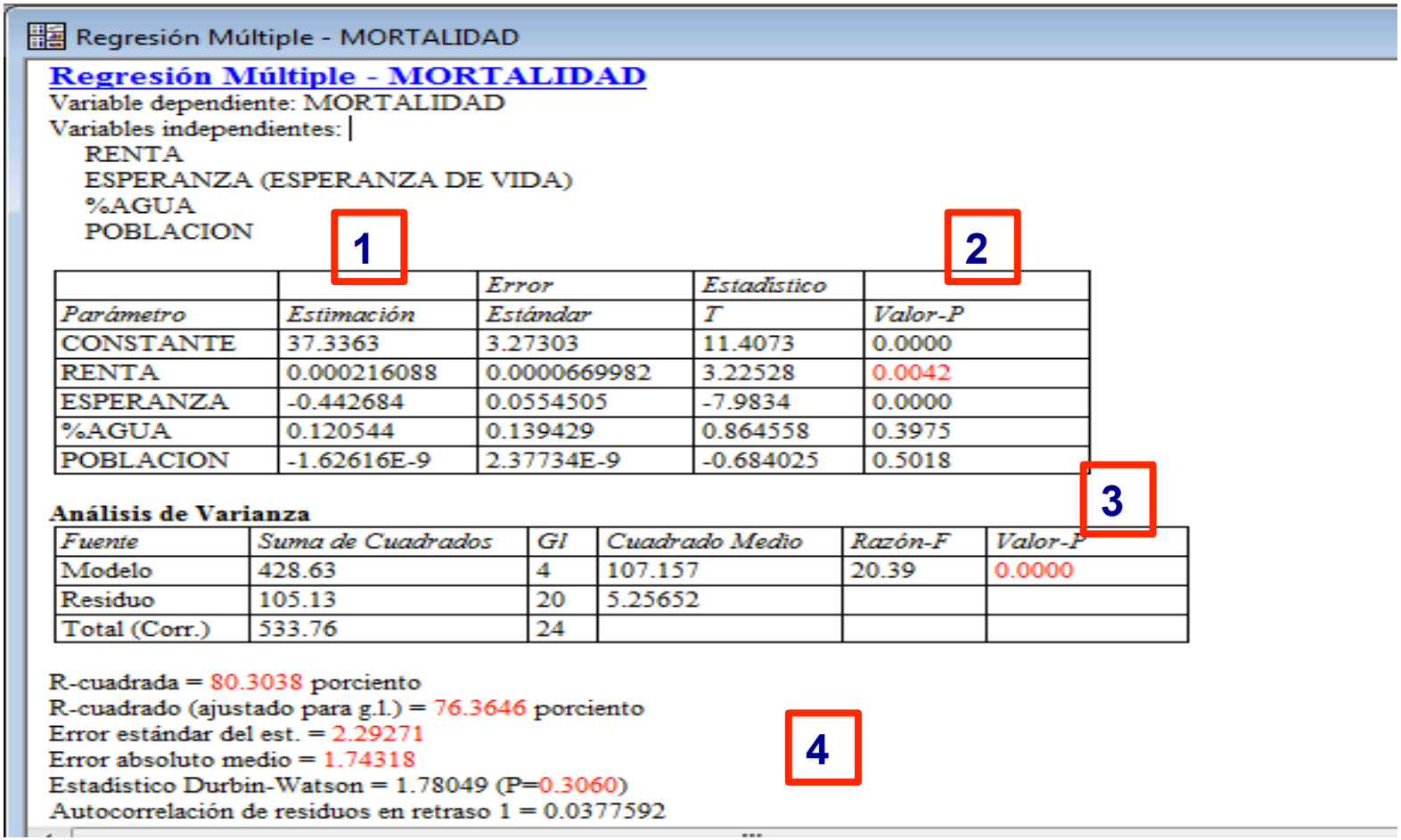
  1. Hay más datos que parámetros desconocidos
  2. n es igual o mayor que k+2
  3. Ninguna de las variables explicativas es combinación lineal exacta de las restantes (colinealidad)

Ejemplo

- Las variables renta per cápita, esperanza de vida, mortalidad, agua (% por habitante) y población se consideran indicadores del nivel de pobreza o riqueza de un país. Se desea conocer el grado de dependencia de la mortalidad en los diferentes países en función del resto de las variables. Los datos se encuentran el el fichero EJEMPLO1(T2).sf3







		Error	Estadistico	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	37.3363	3.27303	11.4073	0.0000
RENTA	0.000216088	0.0000669982	3.22528	0.0042
ESPERANZA	-0.442684	0.0554505	-7.9834	0.0000
%AGUA	0.120544	0.139429	0.864558	0.3975
POBLACION	-1.62616E-9	2.37734E-9	-0.684025	0.5018

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*GI*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	428.63	4	107.157	20.39	0.0000
Residuo	105.13	20	5.25652		
Total (Corr.)	533.76	24			

R-cuadrada = 80.3038 porciento
R-cuadrado (ajustado para g.l.) = 76.3646 porciento
Error estándar del est. = 2.29271
Error absoluto medio = 1.74318
Estadístico Durbin-Watson = 1.78049 (P=0.3060)
Autocorrelación de residuos en retraso 1 = 0.0377592

En el cuadro de Regresión múltiple aparecen las estimaciones del término constante y de los coeficientes de las variables explicativas o predictoras **[1]**, sus errores típicos, el valor de la t-Student para hacer los contrastes individuales sobre las significación de esos parámetros en el modelo ajustado **[2]**. En concreto lo que aparecen son los P-valores para cada contraste

H_0: α_i =0 La variable X_i no es representativa en el modelo
H_0: α_i ≠0 La variable X_i sí es representativa en el modelo

Las variables %AGUA y POBLACION pueden ser eliminadas del modelo

Parámetro	P-valor	Interpretación
α_0 **Constante**	0.0000 < nivel de significación	Rechazamos H_0. El modelo tiene término constante distinto de cero (Test significativo)
α_1 **Renta**	0.0042 < nivel de significación	Rechazamos H_0. La variable RENTA sí es representativa en el modelo (Test significativo)
α_2 **Esperanza**	0.0000 < nivel de significación	Rechazamos H_0. La variable ESPERANZA sí es representativa en el modelo (Test significativo)
α_3 **%Agua**	0.3975 > nivel de significación	NO Rechazamos H_0. La variable %AGUA no es representativa en el modelo (Test NO significativo)
α_4 **Población**	0.5018 > nivel de significación	NO Rechazamos H_0. La variable POBLACION NO es representativa en el modelo (Test NO significativo)

En la tabla de Análisis de Varianza aparece el resultado del contraste de regresión

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$
$$H_1 : \text{existe algún } \alpha_i \neq 0 \text{ para algún i}$$

El P-valor nos indica para qué niveles de significación es significativo (Rechazamos H_0) en conjunto, el modelo de regresión definido en el cuadro primero de Análisis de regresión.

Rechazamos H_0

Análisis de Varianza

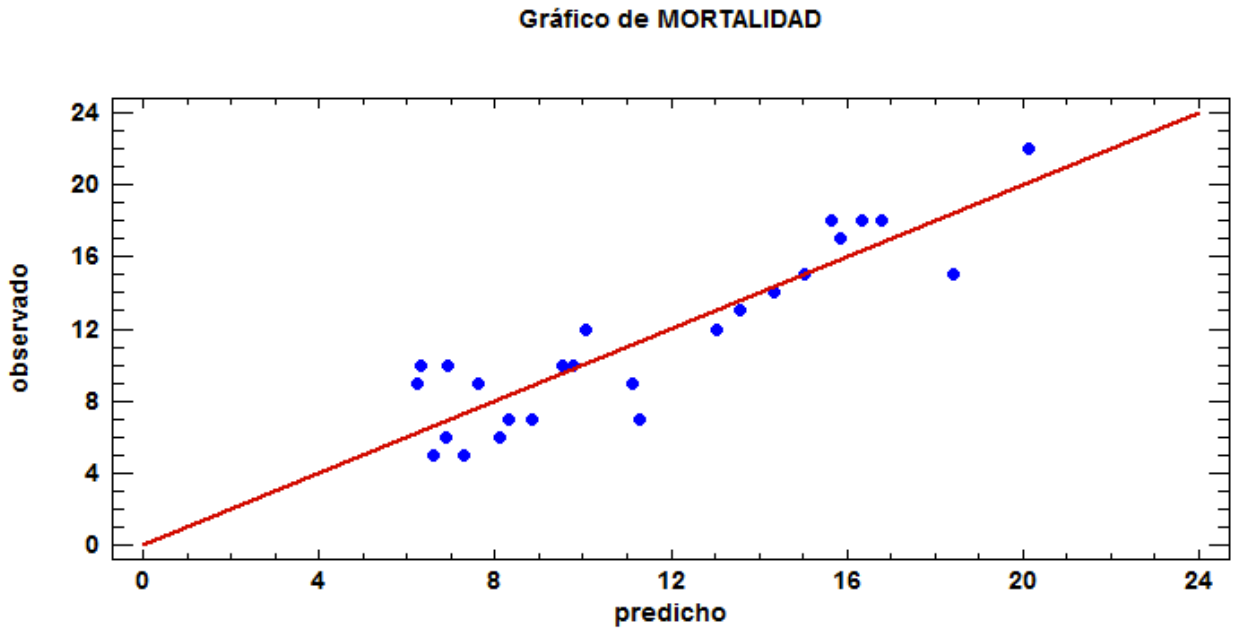
Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	428.63	4	107.157	20.39	0.0000
Residuo	105.13	20	5.25652		
Total (Corr.)	533.76	24			

R-cuadrada = 80.3038 porciento
R-cuadrado (ajustado para g.l.) = 76.3646 porciento
Error estándar del est. = 2.29271
Error absoluto medio = 1.74318
Estadístico Durbin-Watson = 1.78049 (P=0.3060)
Autocorrelación de residuos en retraso 1 = 0.0377592

El modelo tiene un coeficiente de determinación R^2=80.3038% lo que nos indica que no es una mal ajuste.

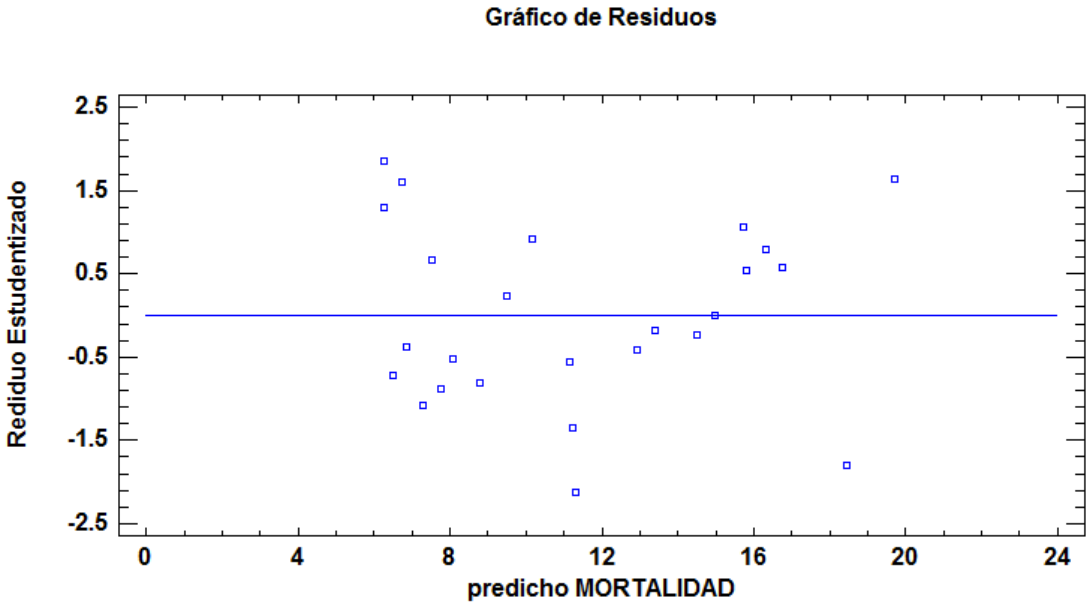
El coeficiente R^2 (ajustado por g.l) es más conveniente para comparar modelos con distinto número de variables independientes.
Para este caso sabemos que las variables %AGUA y POBLACIÓN pueden ser eliminadas del modelo, por lo que podríamos ajustar un modelo con las variables restantes y compararlo con el original mediante este coeficiente ajustado

Gráfico de valores observados frente a predichos



Con este gráfico observamos que el modelo de regresión lineal múltiple propuesto no es el mejor para explicar la relación entre estas variables. El caso ideal es que la nube de puntos coincida con la recta que aparece. Si hay mucha dispersión de los puntos respecto de la recta podemos estar ante un modelo con varianza no constante (heterocedasticidad)

Residuos



Residuos Atípicos

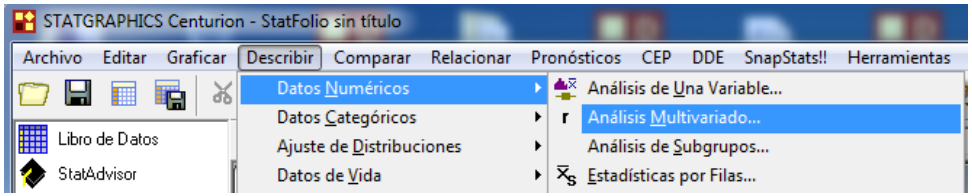
		Y		*Residuo*
Fila	*Y*	*Predicha*	*Residuo*	*Estudentizado*
24	7.0	11.3118	-4.31178	-2.12

El StatAdvisor

La tabla de residuos atípicos enlista todas las observaciones que tienen residuos Estudentizados mayores a 2, en valor absoluto. Los residuos Estudentizados miden cuántas desviaciones estándar se desvía cada valor observado de MORTALIDAD del modelo ajustado, utilizando todos los datos excepto esa observación. En este caso, hay un residuo Estudentizado mayor que 2, pero ninguno mayor que 3.

Muticolinealidad

La estudiamos a través de la matriz de correlaciones. Esta matriz está formada por los coeficientes de correlación para las estimaciones del modelo. Si hay valores próximos a uno, en valor absoluto, fuera de la diagonal principal, es posible que haya multicolinelidad.



Análisis Multivariado

Correlaciones

	RENTA	ESPERANZA	%AGUA	POBLACION
RENTA		0.7094	0.0319	-0.1690
		(25)	(25)	(25)
		0.0001	0.8798	0.4194
ESPERANZA	0.7094		-0.0652	-0.0453
	(25)		(25)	(25)
	0.0001		0.7567	0.8298
%AGUA	0.0319	-0.0652		0.3443
	(25)	(25)		(25)
	0.8798	0.7567		0.0920
POBLACION	-0.1690	-0.0453	0.3443	
	(25)	(25)	(25)	
	0.4194	0.8298	0.0920	

Correlación
(Tamaño de Muestra)
Valor-P

El StatAdvisor

Esta tabla muestra las correlaciones momento producto de Pearson, entre cada par de variables. El rango de estos coeficientes de correlación va de -1 a +1, y miden la fuerza de la relación lineal entre las variables. También se muestra, entre paréntesis, el número de pares de datos utilizados para calcular cada coeficiente. El tercer número en cada bloque de la tabla es un valor-P que prueba la significancia estadística de las correlaciones estimadas. Valores-P abajo de 0.05 indican correlaciones significativamente diferentes de cero, con un nivel de confianza del 95.0%. Los siguientes pares de variables tienen valores-P por debajo de 0.05:
RENTA y ESPERANZA