VIII

REGRESIONES

En capítulos anteriores se describió la forma de determinar si existía una relación significativa entre dos variables cualitativas (Capítulo VI) y cuantitativas (Capítulo VII), es decir, si las variables eran dependientes o independientes. En este capítulo vamos a analizar cómo describir el tipo de función que mejor se ajusta a la posible relación que existe entre variables.

Se define *regresión* como la teoría que trata de expresar mediante una función matemática la relación que existe entre una variable dependiente y una (regresión simple) o varias (regresión múltiple) variables independientes. La obtención de esta función permite predecir cual será el valor de la variable dependiente en función del valor que tome la variable o variables independientes.

La regresión se diferencia de la *correlación* en que esta última estudia el grado de asociación entre las variables, y determina si la relación es o no es significativa, mientras que la regresión, como se mencionó anteriormente, trata de definir la función que mejor explica la relación entre las variables.

En este capítulo vamos a estudiar los modelos de regresión que existen cuando la variable dependiente es cuantitativa y los que hay para variables dependientes cualitativas.

VIII.1. Modelos de regresión para variables dependientes cuantitativas

VIII.1.1. Requisitos

Para aplicar un modelo de regresión entre variables no se requiere que los datos presenten una distribución Normal o que exista homogeneidad de varianzas; sin embargo, para poder determinar si la función obtenida con el modelo de regresión es significativa es necesario aplicar contrastes, y para ello se tienen que cumplir los siguientes requisitos:

1. Los residuos obtenidos del modelo de regresión deben presentar una distribución Normal.

2. Debe existir homocedasticidad en los residuos, es decir, la varianza de los mismos debe ser constante.
3. No debe existir autocorrelación en la serie de residuos (deben ser independientes).
4. En el caso del modelo de regresión múltiple, no debe existir relación lineal entre las variables independientes, es decir, no debe existir multicolinealidad.

VIII.1.2. Regresión simple

Lo primero que hay que hacer cuando se intenta buscar el mejor ajuste entre dos variables, es representar los datos, poniendo en el eje *Y* la variable dependiente y en el eje *X* la independiente. Esto es muy importante, ya que, como se mencionó en el Capítulo I, la representación de los datos es necesaria para ver el tipo de relación que existe entre dos variables y para identificar posibles «outliers».

Las funciones más comunes son las que se muestran en la Figura VIII.1 y se expresan a continuación:

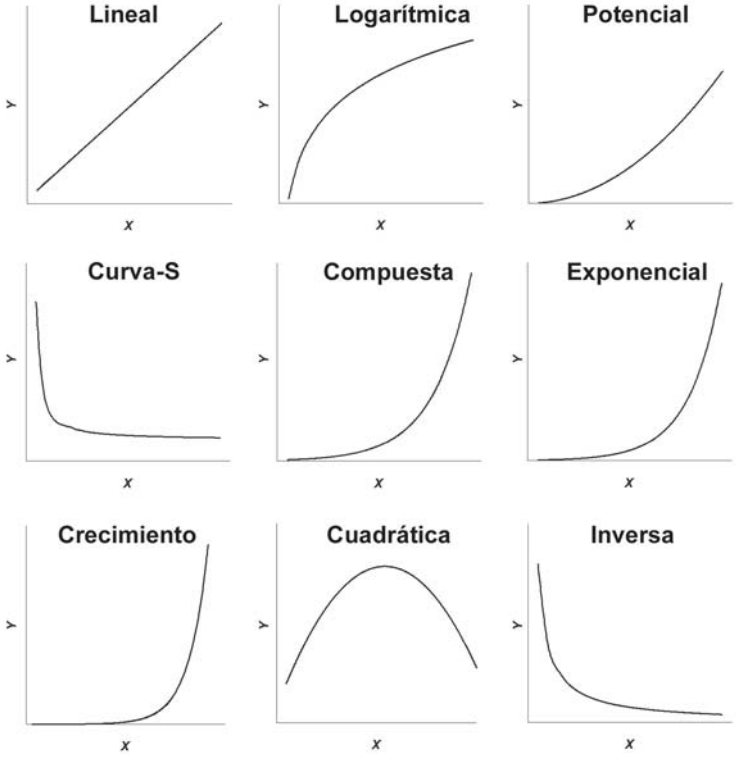


Figura VIII.1. Representación gráfica de las ecuaciones mostradas en el texto.

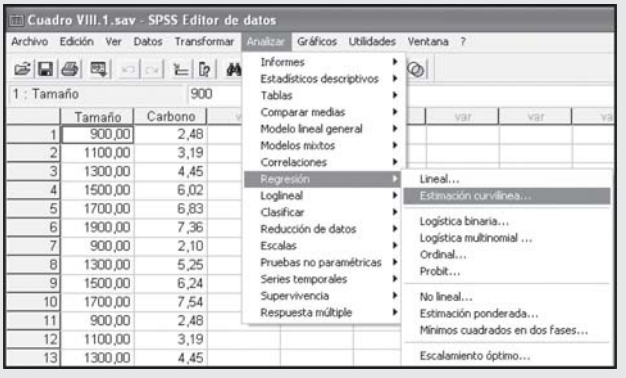
Lineal	$y = a + bx$		
Logarítmica	$y = a + b\ln x$		
Potencial	$y = ax^b$	ó	$\ln y = \ln a + b\ln x$
Exponencial	$y = ae^{bx}$	ó	$\ln y = \ln a + bx$
Compuesta	$y = ab^x$	ó	$\ln y = \ln a + x\ln b$
Curva-S	$y = e^{a+\frac{b}{x}}$	ó	$\ln y = a + \frac{b}{x}$
Cuadrática o Parábola	$y = a + bx + cx^2$		
Crecimiento	$y = e^{(a+bx)}$	ó	$\ln y = a + bx$
Inversa	$y = a + \frac{b}{x}$		

En el Cuadro VIII.1 se muestra la forma de buscar la mejor función que relaciona dos variables.

CUADRO VIII.1. Cálculo de la regresión simple

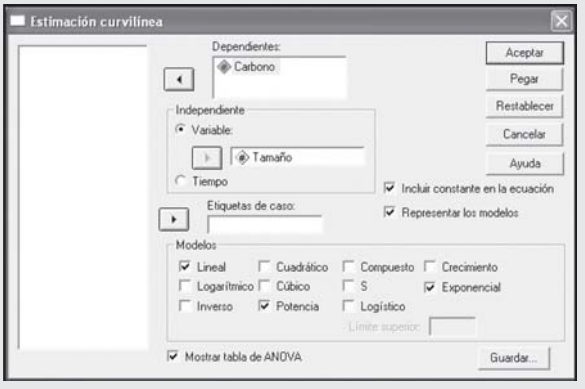
EJEMPLO. En el Archivo **Cuadro VIII.1.sav** se muestra el tamaño de los utrículos (órganos encargados de capturar las presas) de una planta carnívora y su contenido en carbono (µg por utrículo). Se pretende determinar la ecuación que mejor se ajusta a los datos. Utilizaremos el programa SPSS.

Paso 1. Después de introducir nuestras variables, hay que entrar en la sección «Analizar», dentro de esta en «Regresión», y dentro de esta última en «Estimación curvilínea».

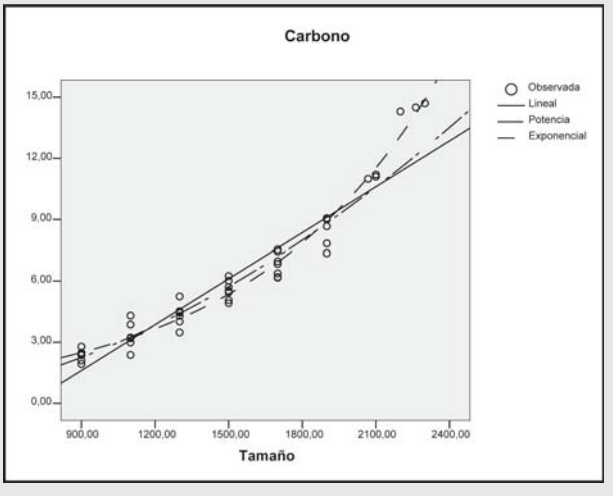


CUADRO VIII.1. *(Continuación)*

Paso 2. Nos saldrá la siguiente ventana en la que ponemos la variable dependiente (*Carbono*) y la independiente (*Tamaño*). Luego indicamos los tipos de funciones que queremos ver en el ajuste. En nuestro caso hemos seleccionado en «Modelos» la «Lineal», «Exponencial» y «Potencia». Es importante que siempre se incluya la constante de la ecuación (el punto de corte de la ecuación por el eje *Y* cuando *X* es cero). Solo en los casos en que sepamos que cuando *X* es cero *Y* también tiene que ser cero, no incluiremos la constante de la ecuación. En nuestro caso lo incluimos y por ello marcamos «Incluir constante en la ecuación». Al marcar «Representar los modelos» en los resultados obtendremos la representación gráfica de los resultados. Por último, es importante marcar «Mostrar tabla de ANOVA» porque obtendremos información sobre el grado de significación del modelo.



Paso 3. Al «Aceptar» se obtiene el gráfico siguiente y los resultados que se muestran a continuación.



CUADRO VIII.1. *(Continuación)*

Para la ecuación lineal:

Resumen del modelo

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,949	,901	,899	1,001

La variable independiente esTamaño.

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	472,192	1	472,192	471,241	,000
Residual	52,105	52	1,002		
Total	524,297	53			

La variable independiente esTamaño.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típico	Beta	t	Sig.
Tamaño	,007	,000	,949	21,708	,000
(Constante)	-5,132	,528		-9,716	,000

Para la ecuación potencial:

Resumen del modelo

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,973	,946	,945	,122

La variable independiente esTamaño.

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	13,682	1	13,682	916,879	,000
Residual	,776	52	,015		
Total	14,458	53			

La variable independiente esTamaño.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típico	Beta	t	Sig.
ln(Tamaño)	1,825	,060	,973	30,280	,000
(Constante)	9,12E-006	,000		2,283	,027

La variable dependiente es ln(Carbono).

CUADRO VIII.1. *(Continuación)*

Y por último, para la ecuación exponencial:

Resumen del modelo

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,973	,947	,946	,121

La variable independiente esTamaño.

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	13,691	1	13,691	927,896	,000
Residual	,767	52	,015		
Total	14,458	53			

La variable independiente esTamaño.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típico	Beta		
Tamaño	,001	,000	,973	30,461	,000
(Constante)	,791	,051		15,602	,000

La variable dependiente es ln(Carbono).

Paso 4. En la interpretación de los resultados es importante tener en cuenta los siguientes criterios:

1. En la tabla de «Resumen del modelo», el «R» (coeficiente de correlación de Pearson) o «R cuadrado» nos indica el grado de correlación que existe entre las variables. El valor de *r* oscila entre -1 y 1, y el valor de r^2 oscila entre 0 y 1. Conforme el valor de r^2 sea más cercano 1, mayor será la correlación y mejor será el ajuste. El r^2 es 0,9 para la lineal y 0,95 para la potencial y exponencial. Por tanto, el grado de ajuste más alto se obtiene usando las ecuaciones exponencial o potencial.

2. En la tabla «ANOVA» obtenemos otro resultado importante, el valor del estadístico *F* y el nivel de significación «Sig.». Como vemos, en los tres modelos la probabilidad es menor de 0,001 y, por tanto, las regresiones son significativas. El valor de *F* es mayor en la exponencial (927,9) que en la potencial (916,8) y la lineal (471,2), siendo los grados de libertad en ambos casos los mismos (1, 52), indicando que la relación exponencial se ajusta mejor a nuestros datos.

3. En la tabla de «Coeficientes» se muestran otros resultados importantes: las estimaciones de las constantes o parámetros que definen la ecuación:

Lineal: $Carbono = -5{,}13 + 0{,}00749 * Tamaño$

Potencial: $Carbono = 9{,}12 * 10^{-6} * Tamaño^{1825}$

Exponencial: $Carbono = 0{,}7906 * e^{0{,}00127 * Tamaño}$

CUADRO VIII.1. *(Continuación)*

4. El último resultado importante a tener en cuenta es el grado de significación asociado a las estimaciones de las constantes de la ecuación, que se muestra en la tabla de «Coeficientes». En el caso de la ecuación lineal y la exponencial, tanto la constante como la pendiente tienen una $p < 0{,}001$, indicando que son significativas. En el caso de la ecuación potencial, también es significativa la pendiente ($p < 0{,}001$) y la constante, pero en este caso con $p = 0{,}027$.

Paso 5. La conclusión es que la ecuación exponencial es la que mejor se ajusta a nuestros datos. La forma de expresar nuestros resultados sería la siguiente: la relación entre tamaño y carbono de los utrículos es significativa ($r^2 = 0{,}95$, $F_{1,52} = 927{,}9$, $p < 0{,}001$) y se ajusta a la ecuación exponencial que se mostró anteriormente.

Sin embargo, para que pueda ser utilizado el estadístico *F* y su nivel de significación, deben cumplirse los requisitos mencionados anteriormente en la Sección VIII.1.1 (homocedasticidad de los residuos, que estos tengan una distribución Normal y que no exista autocorrelación entre ellos). Para comprobarlo es necesario realizar el modelo de regresión lineal.

Paso 6. Como el modelo que mejor se ajusta es la ecuación exponencial, para poder aplicar el modelo de regresión lineal es necesario en primer lugar transformar nuestros datos para que la relación entre ambas variables sea lineal. En la Tabla VIII.1 se muestran las transformaciones que hay que hacer para que la relación entre variables sea lineal.

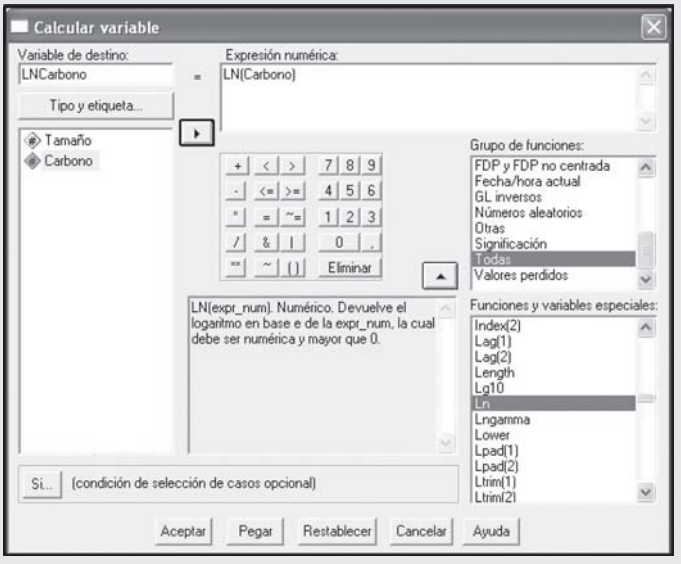
Tabla VIII.1. Transformaciones para conseguir relación lineal.

	Modelo	Transformación *x*	*Transformación y*
Logarítmica	$y = a + b\ln x$	$t(x) = \ln(x)$	$t(y) = y$
Potencial	$y = ax^b$	$t(x) = \ln(x)$	$t(y) = \ln(y)$
Exponencial	$y = ae^{bx}$	$t(x) = x$	$t(y) = \ln(y)$
Curva-S	$y = e^{a+\frac{b}{x}}$	$t(x) = \frac{1}{x}$	$t(y) = \ln(y)$
Inversa	$y = a + \frac{b}{x}$	$t(x) = \frac{1}{x}$	$t(y) = y$

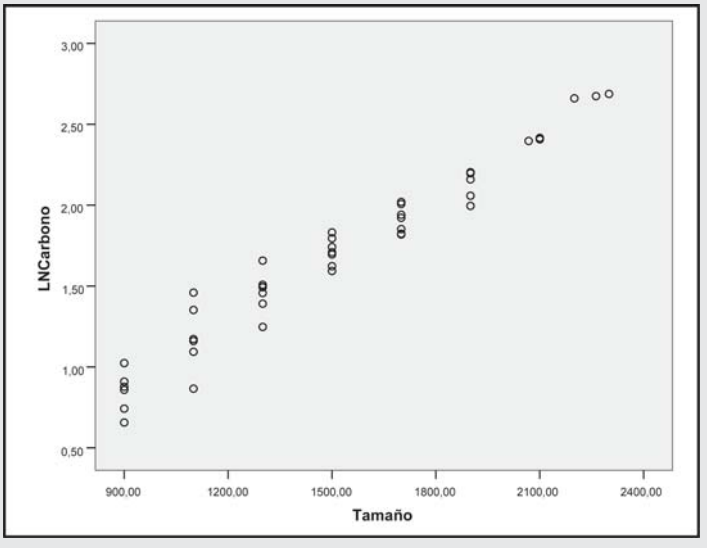
La transformación se realiza en «Transformar» y luego en «Calcular».

CUADRO VIII.1. *(Continuación)*

En la ventana que aparece, en «Variable de destino» creamos una nueva variable *LNCarbono* y en «Expresión numérica» ponemos la transformación que queremos realizar. Como el mejor ajuste lo obteníamos con la exponencial, hay que aplicar el logaritmo a la variable dependiente (*Carbono*) y luego «Aceptar».

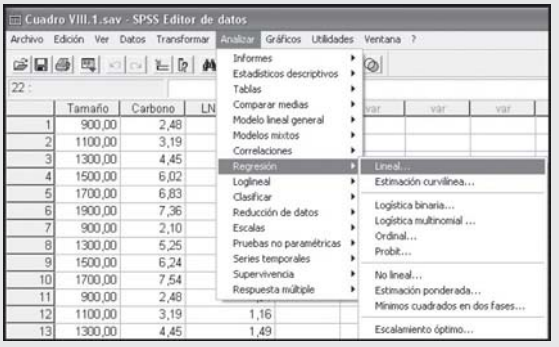


Como se observa en la figura siguiente, es suficiente con transformar la variable dependiente para obtener la relación lineal entre ambas variables.

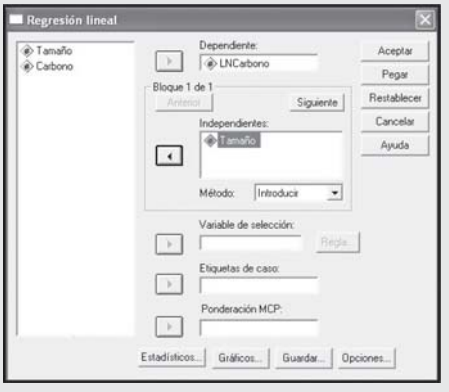


CUADRO VIII.1. *(Continuación)*

Paso 7. Para realizar la regresión lineal, dentro del menú principal nos vamos a «Analizar», luego «Regresión» y, por último, «Lineal».



Paso 8. En la ventana que aparece introducimos como variable dependiente *LNCarbono* y como independiente *Tamaño.*



Paso 9. En la ventana anterior tenemos distintas opciones:

1. En «Método» se puede elegir entre distintas formas de introducir las variables en la ecuación:

- *Introducir*. Nos muestra la ecuación con todas las variables independientes. Es decir, introduce todas las variables en la ecuación aunque no sean significativas.
- *Pasos sucesivos.* En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga una probabilidad para *F* suficientemente pequeña. Las variables ya introducidas en la ecuación de regresión se eliminan de ella si su probabilidad para *F* llega a ser grande. El método termina cuando ya no hay más variables para incluir o excluir. Es decir, solo introduce en la ecuación las variables que son significativas. Es, junto con el anterior, el más usado.

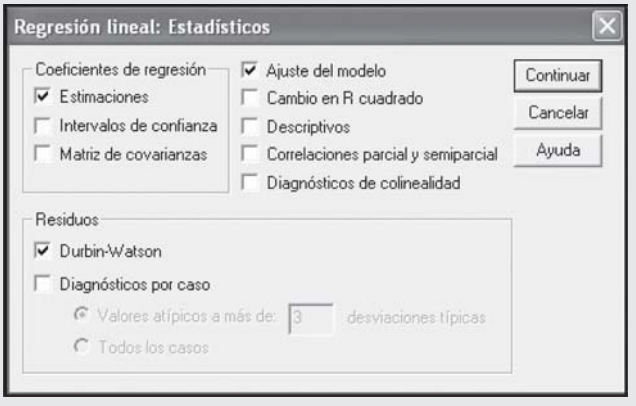
CUADRO VIII.1. *(Continuación)*

- *Eliminar.* Procedimiento para la selección de variables en el que las variables de un bloque se eliminan en un solo paso.
- *Hacia atrás*. Procedimiento de selección de variables en el que se introducen todas las variables en la ecuación y después se van excluyendo una tras otra. Aquella variable que tenga la menor correlación parcial con la variable dependiente será la primera en ser considerada para su exclusión. Si satisface el criterio de eliminación, será eliminada. Tras haber excluido la primera variable, se pondrá a prueba aquella variable, de las que queden en la ecuación, que presente una correlación parcial más pequeña. El procedimiento termina cuando ya no quedan en la ecuación variables que satisfagan el criterio de exclusión.
- *Hacia delante*. Es el proceso contrario al anterior. Va introduciendo primero las variables con mayor significación, hasta que no queda ninguna variable significativa por introducir.

2. En «Variable de selección» se puede introducir una variable que permita seleccionar solo determinados casos para el análisis. Por ejemplo, poniendo 0 y 1 en la variable selección y diciendo que solo se trabaje con los casos que tengan el código 1.

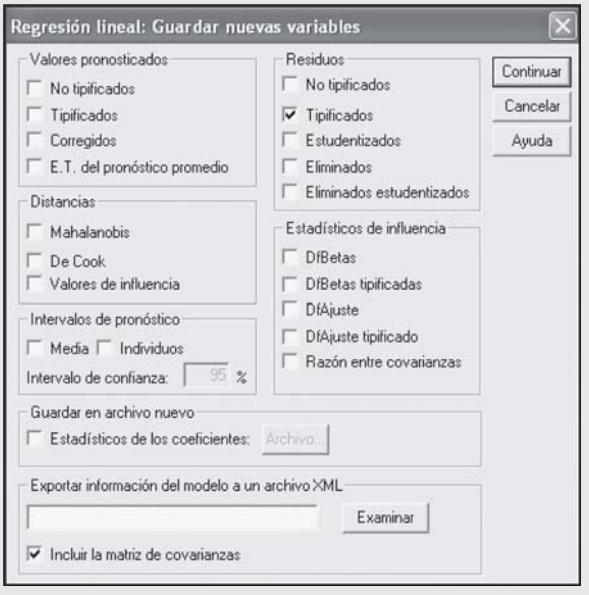
3. «Etiquetas de caso» es otra variable que permite diferenciar los casos a la hora de realizar una representación.

Paso 10. En la ventana del paso 8 entramos en «Estadísticos» y en la ventana que aparece seleccionamos en «Residuos» el test «Durbin-Watson» que permite ver si existe autocorrelación entre los residuos. No tiene sentido ver la colinealidad y seleccionar «Diagnósticos de colinealidad» porque solo hay una variable independiente. Dejamos seleccionado también «Estimaciones» y «Ajuste del modelo».



CUADRO VIII.1. *(Continuación)*

Paso 11. En la ventana del paso 8, en el icono «Guardar» aparece la siguiente ventana donde seleccionamos en «Residuos» los «Tipificados», que son los residuos que usaremos para estudiar la distribución Normal y homocedasticidad.



Paso 12. En la ventana del paso 8, si pulsamos en «Gráficos» nos aparece la siguiente ventana donde podemos especificar que se realice el gráfico entre «ZRESID» (residuos tipificados) y «ZPRED» (valores pronosticados tipificados). Esta gráfica será necesaria para ver la homocedasticidad de los residuos.

CUADRO VIII.1. *(Continuación)*

Paso 13. En la ventana del paso 8, si pulsamos en «Opciones» nos aparece la siguiente ventana donde podemos modificar el nivel de significación para que una variable entre o salga en la regresión por pasos, incluir o no incluir la intersección de la ecuación (lo que se denomina en el programa constante) y excluir casos según alguno de los criterios que se mencionan a continuación:

- *Excluir casos según lista*. Sólo se incluirán en el análisis los casos con valores válidos para todas las variables.

- *Excluir casos según pareja*. Los casos con datos completos para la pareja de variables correlacionadas se utilizan para calcular el coeficiente de correlación en el cual se basa el análisis de regresión. Los grados de libertad se basan en el N mínimo de las parejas.

- *Reemplazar por la media*. Se emplean todos los casos en los cálculos, sustituyendo las observaciones perdidas por la media de la variable.

Paso 14. Por último, en la ventana del paso 8 pulsamos en «Aceptar» para obtener los resultados.

En la tabla «Resumen del modelo» y en la tabla de «ANOVA» se observa cómo el valor de r^2 = 0,95 y el valor de F = 927,9 son iguales a los que se obtenían con la ecuación exponencial (véase paso 4).

CUADRO VIII.1. *(Continuación)*

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,973[a]	,947	,946	,12147	1,256

a. Variables predictoras: (Constante), Tamaño

b. Variable dependiente: LNCarbono

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	13,691	1	13,691	927,896	,000[a]
	Residual	,767	52	,015		
	Total	14,458	53			

a. Variables predictoras: (Constante), Tamaño

b. Variable dependiente: LNCarbono

En la tabla de «Coeficientes», que se muestra a continuación, se observa que, tanto la constante de la ecuación (-0,235) como la pendiente (0,001) son significativas con una $p \leq 0{,}001$.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-,235	,064		-3,664	,001
	Tamaño	,001	,000	,973	30,461	,000

a. Variable dependiente: LNCarbono

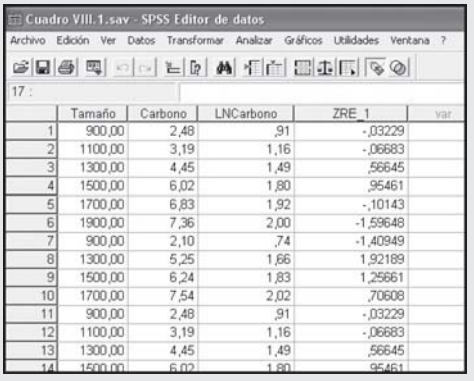
La forma de expresar nuestros resultados sería la siguiente: La relación entre tamaño y carbono de los utrículos es significativa (r^2 = 0,95, $F_{1,52}$ = 927,9, p < 0,001) y la ecuación que relaciona ambas variables es la siguiente:

$$\ln(Carbono) = - 0{,}235 + 0{,}001 * Tamaño$$

Ahora que ya hemos conseguido la linealidad entre las variables y hemos realizado la regresión lineal podemos ver si es correcto usar el estadístico *F* y su nivel de significación, examinando los requisitos de distribución Normal y homocedasticidad de los residuos (diferencia entre el valor observado de la variable dependiente y el valor ajustado por la ecuación) y que no exista autocorrelación entre ellos.

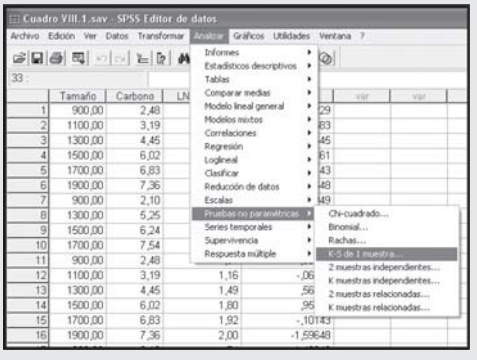
CUADRO VIII.1. *(Continuación)*

Paso 15. Distribución Normal de los residuos. Volviendo a la página principal veremos que se ha creado una nueva variable que contiene los residuos tipificados (ZRE_1).

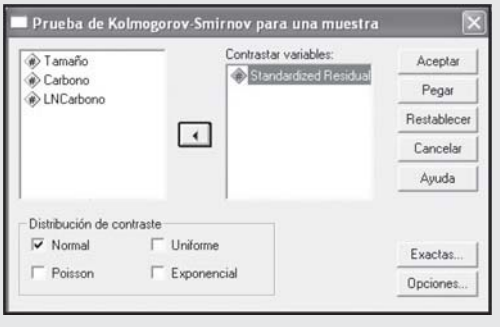


	Tamaño	Carbono	LNCarbono	ZRE_1	var
1	900,00	2,48	,91	-,03229	
2	1100,00	3,19	1,16	-,06683	
3	1300,00	4,45	1,49	,56645	
4	1500,00	6,02	1,80	,95461	
5	1700,00	6,83	1,92	-,10143	
6	1900,00	7,36	2,00	-1,59648	
7	900,00	2,10	,74	-1,40949	
8	1300,00	5,25	1,66	1,92189	
9	1500,00	6,24	1,83	1,25661	
10	1700,00	7,54	2,02	,70608	
11	900,00	2,48	,91	-,03229	
12	1100,00	3,19	1,16	-,06683	
13	1300,00	4,45	1,49	,56645	
14	1500,00	6,02	1,80	,95461	

Para ver la distribución Normal de los residuos entramos en «Analizar», luego en «Pruebas no paramétricas» y, por último, en «KS de 1 muestra», para hacer el test de Kolmogorov-Smirnov.



Nos aparece la siguiente ventana en la cual, dentro de «Contrastar variables», introducimos la variable de la que queremos ver si tiene una distribución Normal, en este caso «Standardized Residual».



CUADRO VIII.1. *(Continuación)*

Al «Aceptar» nos aparece la siguiente tabla de resultados «Prueba de Kolmogorov-Smirnov para una muestra» en la que se observa que p = 0,665 y, por lo tanto, se cumple la hipótesis nula de que los residuos se ajustan a una distribución Normal.

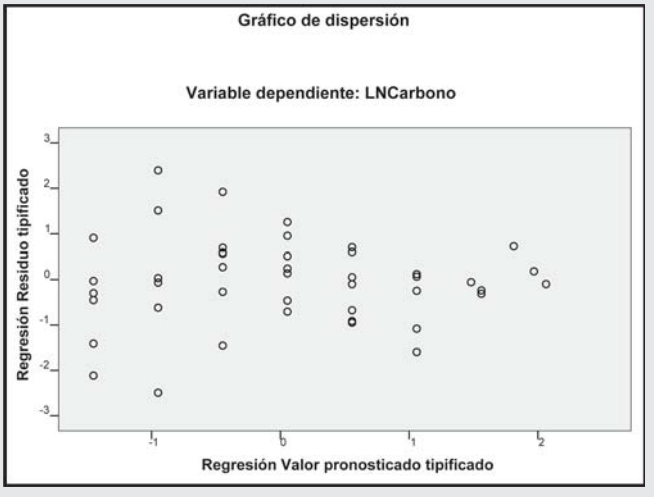
Prueba de Kolmogorov-Smirnov para una muestra

		Standardized Residual
N		54
Parámetros normales [a,b]	Media	,0000000
	Desviación típica	,99052111
Diferencias más extremas	Absoluta	,099
	Positiva	,066
	Negativa	-,099
Z de Kolmogorov-Smirnov		,728
Sig. asintót. (bilateral)		,665

a. La distribución de contraste es la Normal.

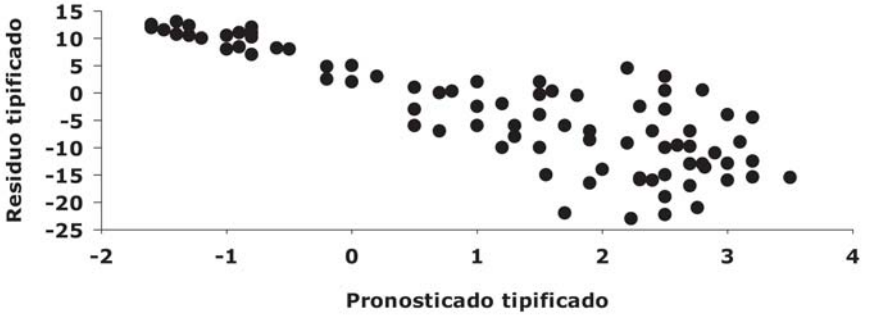
b. Se han calculado a partir de los datos.

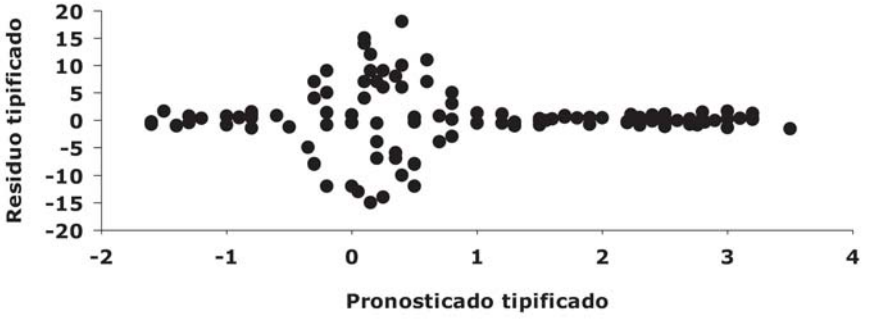
Paso 16. Homocedasticidad de los residuos. Uno de los resultados obtenidos es el gráfico que se muestra a continuación (y que habíamos especificado en el paso 12) entre los residuos tipificados y las predicciones. Para que exista estabilidad en la variabilidad de los residuos, esta debe ser más o menos constante, no aumentando ni disminuyendo en los extremos o el centro de la gráfica. Vemos cómo en este caso los residuos varían de forma similar a lo largo de todo el recorrido y, por tanto, aceptamos que existe homocedasticidad en los residuos**.**

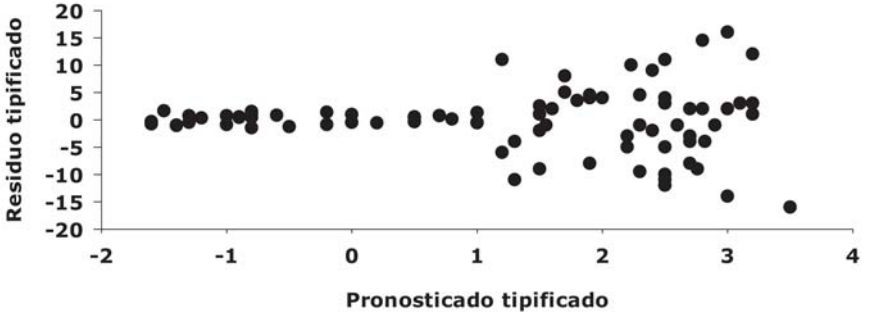


CUADRO VIII.1. *(Continuación)*

Las gráficas que se muestran a continuación son ejemplos en los que no se cumple la homocedasticidad de los residuos, porque su variabilidad no es constante, sino que cambia dependiendo de los valores pronosticados. En la primera gráfica se observa, además, que el ajuste no es lineal, ya que no se mantiene la tendencia horizontal.







CUADRO VIII.1. *(Continuación)*

Paso 17. Autocorrelación de los residuos. Para determinar si no existe autocorrelación en los residuos se utiliza el contraste de Durbin-Watson. En el paso 14, en la tabla «Resumen del modelo», se ve el valor del contraste $\hat{d}$ = 1,256. Las tablas de Durbin-Watson (Tabla 7 del Apéndice o Archivo **Tabla 7.doc**) para n = 54 y α = 0,05 proporcionan los valores aproximados de d_L = 1,527 y d_U = 1,601.

- Si $0 < \hat{d} < d_L$ se rechaza H_0 y aceptamos la existencia de autocorrelación positiva.
- Si $d_L < \hat{d} < d_U$ el contraste no es concluyente.
- Si $d_U < \hat{d} < 4 - d_U$ se acepta H_0 y no hay autocorrelación.
- Si $4 - d_U < \hat{d} < 4 - d_L$ el contraste no es concluyente.
- Si $4 - d_L < \hat{d} < 4$ se rechaza H_0 y aceptamos la existencia de autocorrelación negativa.

Por lo tanto, como $\hat{d}$ (1,256) < d_L (1,527) se rechaza H_0 y existe autocorrelación positiva entre los residuos. La autocorrelación aparece generalmente porque las medidas se toman a lo largo del tiempo y los errores se deben a los procedimientos utilizados. Una posible solución consiste en aleatorizar el proceso de medida en la fase de toma de datos.

Paso 18. En conclusión, aunque el modelo de regresión es muy significativo, no se debe utilizar el estadístico *F* ni el grado de significación que se obtiene, porque existe autocorrelación (los residuos no son independientes).

VIII.1.3. Regresión múltiple lineal

Hasta ahora hemos visto regresiones en las que había una sola variable independiente. Sin embargo, es muy frecuente que necesitemos estudiar si nuestra variable dependiente está relacionada con más de una variable. En este caso es necesario usar la regresión múltiple.

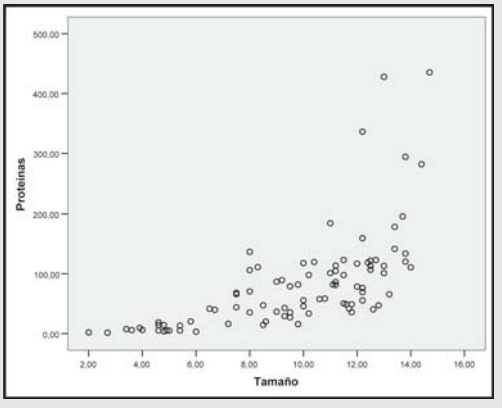
Al igual que ocurría con la regresión simple, las relaciones entre variables pueden ser lineales o no lineales. En este apartado vamos a tratar solo con aquellas situaciones en las que, aunque algunas variables no tengan relación lineal, es posible obtener la linealidad mediante alguna transformación. En el siguiente apartado se explicará la forma de obtener la regresión cuando la relación no es lineal.

En el Cuadro VIII.2 se muestra un ejemplo de regresión múltiple lineal.

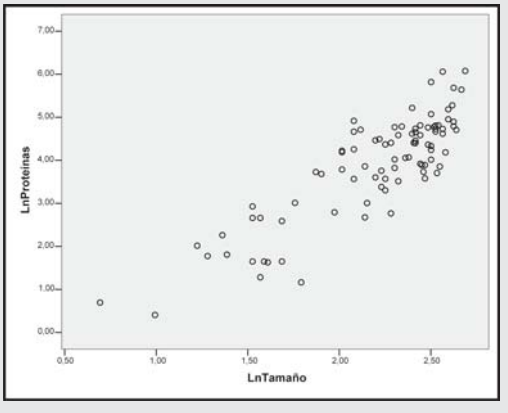
CUADRO VIII.2. Regresión múltiple lineal

EJEMPLO. En el Archivo **Cuadro VIII.2.sav** se muestran datos sobre el tamaño (en mm) de la larva y el contenido en proteínas de la larva (en µg por larva) de una especie de pez recolectadas en distintas estaciones de muestreo donde se midió la temperatura (º C) y la concentración de oxígeno del agua (mg ml^{-1}). Se quiere determinar si el contenido en proteínas de la larva depende del tamaño de la larva, concentración de oxígeno y/o de la temperatura, y encontrar la función que los relaciona.

Paso 1. Por medio de los pasos descritos en el Cuadro VIII.1 o simplemente representando en una gráfica, se observa que la relación entre el tamaño de la larva y su contenido en proteínas no es lineal. Con las otras dos variables, temperatura y concentración de oxígeno, no se observa que la relación no sea lineal.

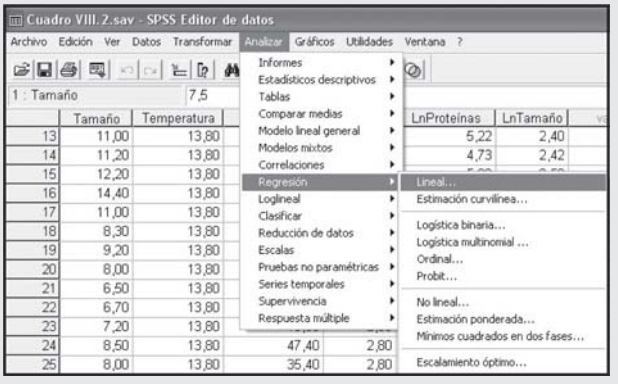


En esta situación es necesario primero hacer que las relaciones sean lineales y, para ello, como la relación entre proteína y tamaño de la larva es potencial (Cuadro VIII.1 para ver los pasos a seguir), aplicamos el logaritmo a las dos variables. Como se ve en el nuevo gráfico, la relación se hace lineal.

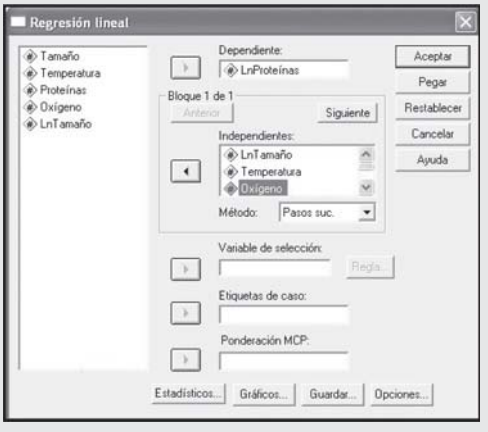


CUADRO VIII.2. *(Continuación)*

Paso 2. En el programa SPSS entramos en «Analizar», luego «Regresión» y, por último «Lineal...».



Paso 3. En la ventana que aparece introducimos las variables transformadas con los logaritmos. La dependiente es el contenido en proteínas de la larva y las independientes el tamaño de la larva, concentración de oxígeno y la temperatura. Se pueden hacer pruebas, introduciendo como variables independientes los datos transformados logarítmicamente o sin transformar, para obtener la mejor relación y que cumpla con todos los requisitos.



Las distintas opciones que aparecen en la ventana se explicaron en los pasos del 9 al 13 del Cuadro VIII.1, con la excepción de que en este caso se selecciona «Diagnósticos de colinealidad» (paso 10 del Cuadro VIII.1) para ver si existe relación entre las variables independientes. En «Método» utilizamos «Pasos suc.» para que el modelo solo incluya las variables independientes que son significativas.

CUADRO VIII.2. *(Continuación)*

Paso 4. Los resultados del análisis son los siguientes, de los cuales la mayoría de los parámetros ya se explicaron en el Cuadro VIII.1.

1. En la tabla «Resumen del modelo» el r^2 del segundo modelo (0,81) es mayor que el del primer modelo (0,74) y, por tanto, el ajuste es mejor (siempre será así cuando se añaden variables a un modelo).

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,858[a]	,737	,734	,62496
2	,900[b]	,810	,805	,53421

a. Variables predictoras: (Constante), LnTamaño

b. Variables predictoras: (Constante), LnTamaño, Temperatura

2. En la tabla «ANOVA» se muestra que ambos modelos, con una y dos variables independientes, son significativos, con $p < 0{,}001$.

ANOVA[c]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	96,085	1	96,085	246,011	,000[a]
	Residual	34,370	88	,391		
	Total	130,455	89			
2	Regresión	105,627	2	52,813	185,064	,000[b]
	Residual	24,828	87	,285		
	Total	130,455	89			

a. Variables predictoras: (Constante), LnTamaño

b. Variables predictoras: (Constante), LnTamaño, Temperatura

c. Variable dependiente: LnProteínas

3. En la tabla «Coeficientes» vemos que el primer modelo solo incluye como variable independiente el tamaño de la larva. El segundo incluye también la temperatura. No aparece un tercer modelo porque la concentración de oxígeno no es una variable significativa que esté relacionada con las proteínas de la larva. En el segundo modelo, tanto la intersección de la ecuación (la constante), como las dos variables independientes, son significativas con una $p < 0{,}001$.

CUADRO VIII.2. *(Continuación)*

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	-1,567	,354		-4,434	,000		
	LnTamaño	2,481	,158	,858	15,685	,000	1,000	1,000
2	(Constante)	3,085	,859		3,589	,001		
	LnTamaño	2,653	,138	,918	19,165	,000	,953	1,049
	Temperatura	-,339	,059	-,277	-5,782	,000	,953	1,049

a. Variable dependiente: LnProteínas

En resumen, como el segundo modelo explica una proporción claramente mayor de la varianza observada en las proteínas de la larva (81%), es significativo con $p < 0{,}001$, y, tanto la constante como las dos variables independientes son significativas con una $p < 0{,}001$, elegiremos este segundo modelo en vez del primero.

La forma de expresar nuestros resultados sería la siguiente: existe una relación significativa entre las proteínas de la larva con el tamaño de la larva y la temperatura ($r^2 = 0{,}81$, $F_{2,87} = 185{,}06$, $p < 0{,}001$), que se describe por medio de la siguiente ecuación:

$$\ln(\textit{Proteínas}) = 3{,}085 + 2{,}653*\ln(\textit{Tamaño}) - 0{,}339*\textit{Temperatura}$$

Ahora podemos ver si es correcto usar el estadístico *F* y su nivel de significación, examinando los requisitos de distribución Normal y homocedasticidad de los residuos, que no exista autocorrelación entre ellos y tampoco colinealidad entre las variables independientes.

Paso 5. Distribución Normal de los residuos. Se sigue el proceso descrito en el paso 15 del Cuadro VIII.1.

Prueba de Kolmogorov-Smirnov para una muestra

		Standardized Residual
N		90
Parámetros normales [a,b]	Media	,0000000
	Desviación típica	,98870020
Diferencias más extremas	Absoluta	,077
	Positiva	,048
	Negativa	-,077
Z de Kolmogorov-Smirnov		,730
Sig. asintót. (bilateral)		,660

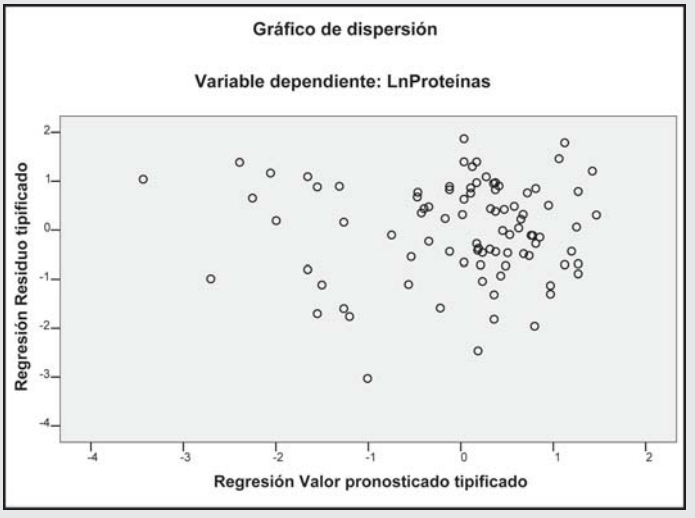
a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Se observa que $p = 0{,}66$, por tanto, se acepta la hipótesis nula de que los residuos presentan una distribución Normal.

CUADRO VIII.2. *(Continuación)*

Paso 6. Homocedasticidad de los residuos. Se sigue el proceso descrito en el paso 16 del Cuadro VIII.1. Como se observa en la gráfica siguiente, la variabilidad de los residuos se mantiene más o menos constante a lo largo de todo el rango de los valores pronosticados tipificados. Por tanto, aceptamos que existe homocedasticidad en los residuos.



Paso 7. Autocorrelación de los residuos. Se sigue el proceso descrito en el paso 17 del Cuadro VIII.1. En la tabla «Resumen del modelo», se ve el valor del contraste $\hat{d}$ = 1,788. Las tablas de Durbin-Watson (Tabla 7 del Apéndice o Archivo **Tabla 7.doc**) para n = 90 y α = 0,05 proporcionan los valores de d_L = 1,612 y d_U = 1,703 (considerando dos variables independientes).

- Si $0 < \hat{d} < d_L$ se rechaza H_0 y aceptamos la existencia de autocorrelación positiva.
- Si $d_L < \hat{d} < d_U$ el contraste no es concluyente.
- Si $d_U < \hat{d} < 4 - d_U$ se acepta H_0 y no hay autocorrelación.
- Si $4 - d_U < \hat{d} < 4 - d_L$ el contraste no es concluyente.
- Si $4 - d_L < \hat{d} < 4$ se rechaza H_0 y aceptamos la existencia de autocorrelación negativa.

Por lo tanto, como d_L (1,612) < $\hat{d}$ (1,788) < $4 - d_U$ (2,297) se acepta H_0 y no hay autocorrelación entre los residuos.

CUADRO VIII.2. *(Continuación)*

Paso 8. Colinealidad. En la tabla «Coeficientes» presentada en el paso 4 se muestran dos estadísticos denominados «Tolerancia» y «FIV» (Factor de Inflación de la Varianza, el inverso de la tolerancia) que se utilizan para estudiar la colinealidad. La tolerancia es la proporción de la varianza de cada variable independiente que no es explicada por las restantes variables independientes, y se obtiene restando de la unidad el coeficiente de determinación r^2 de una regresión múltiple realizada con cada una de las variables explicativas como dependiente de las restantes. Aunque no existen reglas fijas, una tolerancia muy pequeña (por ejemplo inferior al 10% (menor de 0,1), o FIV > 10) muestra una variable que casi es combinación lineal de las restantes, lo que indicaría un posible problema de colinealidad. En nuestro ejemplo la tolerancia es muy alta (0,953) y FIV bajo (1,049), por lo que no existe problema.

En la tabla siguiente «Diagnósticos de colinealidad» se muestran otros resultados para estudiar este problema. En ellos se muestran el «Autovalor», que indica las dimensiones subyacentes: si hay varios autovalores muy próximos a cero, ello indicaría la presencia de colinealidad. Para su mejor valoración se calcula el «Indice de condición», o raíz cuadrada del cociente entre el mayor autovalor y los restantes. Un índice de condición mayor que 15 indica la posible presencia de colinealidad, y si es mayor que 30 indica un serio problema. Sin embargo, dado que la constante del modelo añade una dimensión más, debe matizarse con las «Proporciones de la varianza» de cada coeficiente de regresión que explica cada dimensión (en general cada dimensión explica un porcentaje alto de solamente una variable explicativa). Si una dimensión con índice de condición elevado explica un porcentaje alto de dos o más variables explicativas (la constante no cuenta), existe un serio problema de colinealidad. En nuestro ejemplo no hay colinealidad, ya que una dimensión (la tercera) tiene índice de condición elevado (36,903), pero solo explica una proporción importante de la variable temperatura (y de la constante, pero ésta no cuenta). Cuando existe problema de colinealidad, la mejor solución puede ser prescindir de alguna variable explicativa que esté muy correlacionada con las demás.

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza		
				(Constante)	LnTamaño	Temperatura
1	1	1,982	1,000	,01	,01	
	2	,018	10,639	,99	,99	
2	1	2,976	1,000	,00	,00	,00
	2	,022	11,559	,03	,99	,03
	3	,002	36,903	,97	,00	,97

a. Variable dependiente: LnProteínas

Paso 9. En conclusión, se cumplen razonablemente las hipótesis del modelo de regresión.

VIII.1.4. Otras regresiones simples o múltiples no lineales

Muchas veces las variables se ajustan a un tipo de regresión que no es ninguna de las comunes que vienen descritas en la mayoría de los programas estadísticos y que se mostraron en el Apartado VIII.1.2. Sin embargo, algunos programas estadísticos permiten ver si la relación entre una variable dependiente y una o varias variables independientes se ajusta a una determinada ecuación. Para ello, es necesario tener una idea previa de la ecuación que mejor se puede ajustar a la tendencia que observamos entre las variables. En este apartado se describirán algunas de las relaciones no lineales que se observan más frecuentemente entre distintas variables.

VIII.1.4.1. Curva logística

Muchos fenómenos en la naturaleza se ajustan a una curva logística como por ejemplo los cambios en la abundancia de una población en el tiempo (Smith & Smith 2000), el avance científico a lo largo del tiempo dentro de una determinada línea de investigación (Solla Price, *Little Science, Big Science*, 1963 -citado en Callon y col. 1995), el ciclo de vida de uso que se le da a una línea de ferrocarril (Ingleda & Coto 2003) y un largo etc. (Figura VIII.2).

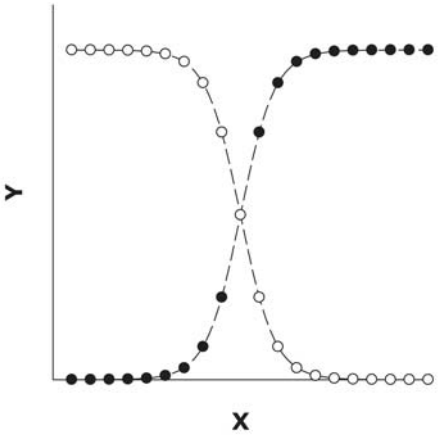


Figura VIII.2. Curva logística con pendiente negativa (○) y positiva (●).

La curva logística viene definida por la siguiente ecuación:

$$Y = \frac{a}{1 + e^{(b-cX)}}$$

La constante *a* delimita el límite superior de la curva y es igual a *K* o capacidad de carga de la población (número máximo de individuos de esa población) en el caso de que se modele a una población (Figura VIII.2.). La constante *c* determina

la pendiente, y en el caso de poblaciones es igual a *r* o tasa de crecimiento de la población. La constante *b* define el tamaño de la llamada fase de latencia en el caso de poblaciones, la fase con valores más pequeños antes de iniciar la pendiente hasta alcanzar el valor máximo.

VIII.1.4.2. Curva de crecimiento de von Bertalanffy

La curva de crecimiento de muchos organismos se ajusta bien a la ecuación de von Bertalanffy (Figura VIII.3):

$$L_t = L_\infty \left(1 - \exp^{-k(t-t_0)}\right)$$

donde L_t es la longitud del individuo a la edad t, L_∞ es la longitud máxima que alcanza el individuo cuando el crecimiento cesa, k es la constante de crecimiento expresada en tiempo^{-1} y t_0 es la edad hipotética que tendría un individuo que tuviese un tamaño cero. Por lo tanto, la variable dependiente sería L_t, la variable independiente la edad del individuo (t) y las constantes de esta ecuación serían k, L_∞ y t_0.

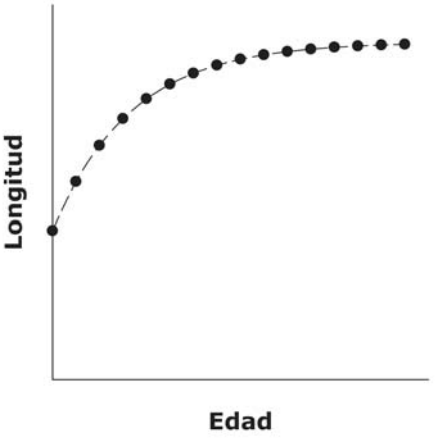


Figura VIII.3. Curva de crecimiento de von Bertalanffy.

VIII.1.4.3. Curva de crecimiento de Gompertz

En algunas especies, como por ejemplo los peces, crustáceos y moluscos, en las primeras fases del ciclo de vida (fase larvaria) el crecimiento puede ajustarse mejor a la ecuación de Gompertz (Hernandez-Llamas & Ratkowsky 2004) (Figura VIII.4):

$$L_t = ae^{(-be^{(-ct)})}$$

donde L_t es la longitud del individuo a la edad t, a es una constante que representa el valor máximo asintótico de la curva, t es la edad del individuo y, b y c son las otras dos constantes de la ecuación.

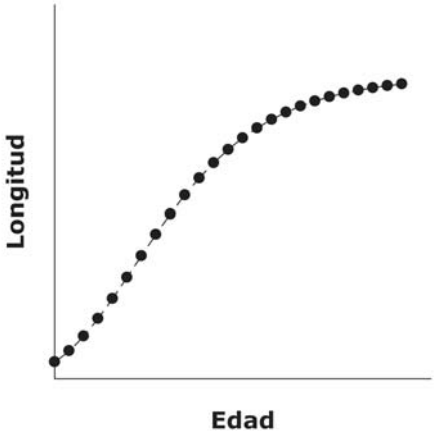


Figura VIII.4. Curva de crecimiento de Gompertz.

VIII.1.4.4. Relación entre tasas y variables

La tasa de crecimiento de una población bacteriana, de levaduras o de una fitoplanctónica a distintas concentraciones de nutrientes o de temperatura, la tasa fotosintética de plantas a diferentes intensidades de luz, entre otros muchos ejemplos, se ajustan bien a la ecuación de Monod (Figura VIII.5):

$$\mu = \frac{\mu_{\max} P}{K_s + P}$$

donde μ es la tasa expresada en tiempo^{-1} y P es la variable (luz, temperatura, concentración de nutrientes, etc) que provoca la variación de la tasa (de crecimiento, fotosintética, etc).

Las constantes de la ecuación son μ_{max} (en tiempo^{-1}), que es la tasa máxima asintótica de la ecuación y K_s que indica el valor de la variable P cuando se alcanza la mitad de la tasa máxima.

La ecuación de Monod presenta el problema de que la tasa solo alcanza el valor cero cuando el valor de la variable también es cero. Sin embargo, se observa frecuentemente que la tasa puede hacerse cero antes de que el valor de la variable explicativa sea cero. Por ejemplo, la tasa de crecimiento poblacional del fitoplancton se suele hacer cero a concentraciones de nutrientes (fosfato, nitrógeno o silicato) superiores a cero (Frangópulos y col. 2004). Esta concentración mínima de nutrientes a la que la tasa de crecimiento es cero, tiene gran importancia en

términos de competencia entre las especies, porque aquellas especies para las que esta concentración es mayor significa que compiten peor por ese nutriente. La ecuación siguiente corrige ese problema de la ecuación de Monod, ya que contempla la posibilidad de que la tasa se haga cero antes de que el parámetro llegue a cero (Figura VIII.5):

$$\mu = \mu_{\max}\left(1 - e^{(-b(P - K_{\min}))}\right)$$

donde K_{min} es el valor del parámetro al cual la tasa es cero. Como se mencionó anteriormente, en los estudios de competencia, a medida que K_{min} es mayor, la especie es peor competidora por el recurso (la variable *P* de la ecuación).

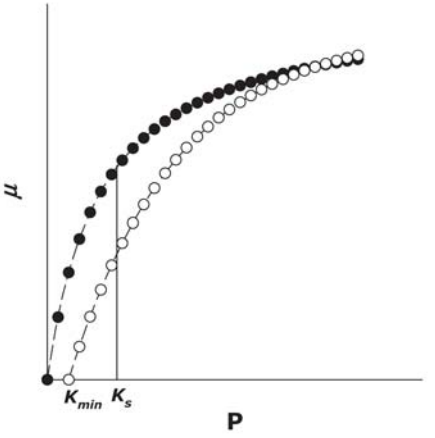


Figura VIII.5. Ecuación de Monod (●) y ecuación que incluye K_{min} (○).

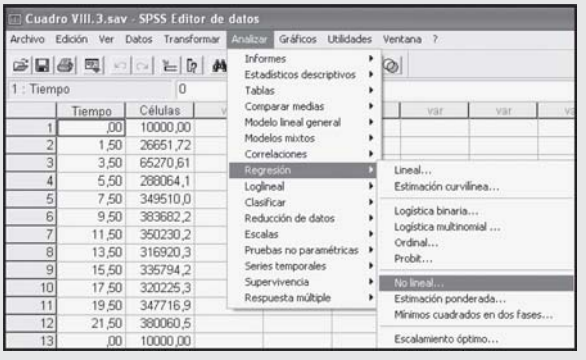
Esta ecuación que incluye K_{min} admite que se incluyan tasas negativas, lo cual también es frecuente observarlo en la naturaleza. Por ejemplo, que el crecimiento de una población bacteriana sea negativo y, por tanto, se reduzca la abundancia de la población cuando la concentración de nutrientes sea muy baja. En el caso de la ecuación de Monod no existe la posibilidad de incluir tasas negativas.

En el Cuadro VIII.3 se describe la forma de obtener cualquier tipo de regresión entre variables y si la relación es significativa.

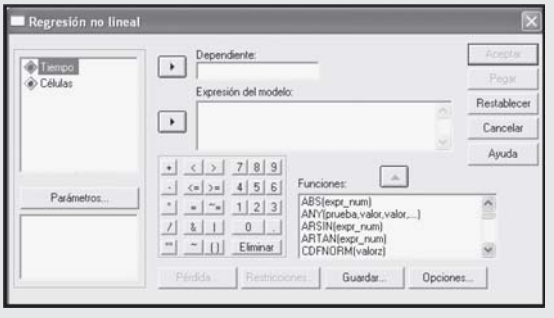
CUADRO VIII.3. Cálculo de regresiones no lineales con el programa SPSS

EJEMPLO. El crecimiento de una población de fitoplancton sigue normalmente una curva de crecimiento logístico. En el archivo **Cuadro VIII.3.sav** se muestra la abundancia de fitoplancton (células ml^{-1}) en cultivo a lo largo del tiempo (en días), teniendo varias réplicas para cada tiempo. Se quiere estimar la curva logística de crecimiento que mejor se ajusta a esos datos.

Paso 1. Después de introducir nuestras variables, hay que entrar en la sección «Analizar», dentro de esta en «Regresión», y dentro de esta última en «No lineal…».

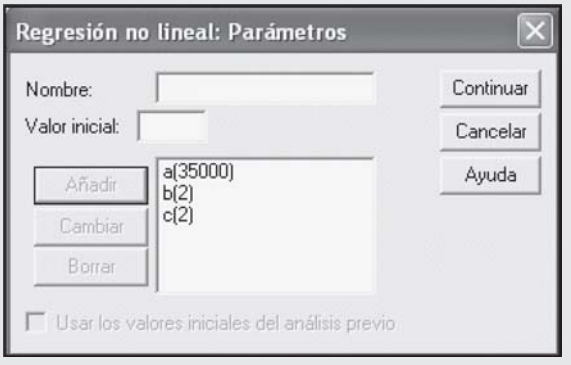


Paso 2. Nos aparecerá el siguiente cuadro en el cual es necesario, en primer lugar, definir los valores iniciales que le damos a las constantes de la ecuación y, para ello entramos en «Parámetros».



CUADRO VIII.3 *(Continuación)*

Paso 3. Nos aparece el siguiente cuadro, donde hay que introducir el nombre de la constante en «Nombre» y el valor de la misma en «Valor inicial». En nuestro caso será $a = 35000$, $b = 2$ y $c = 2$. Es importante introducir valores cercanos a lo que puede ser el resultado final, ya que si no se corre el riesgo de que el programa no sea capaz de ajustar la ecuación.



Paso 4. Una vez introducidas las constantes, el paso siguiente es introducir en «Expresión del modelo» la ecuación que deseamos ajustar a nuestras variables, como se muestra en la siguiente ventana. Es importante mencionar que es posible introducir varias variables independientes, es decir, es posible realizar una regresión múltiple.

CUADRO VIII.3 *(Continuación)*

Paso 5. En los resultados nos aparece la tabla «Historial de iteraciones», con las iteraciones que el programa ha realizado hasta llegar al cálculo de los parámetros.

Historial de iteraciones[b]

Número de iteraciones [a]	Suma de cuadrados residual	Parámetro		
		a	b	c
1.0	3,272E+012	35000,000	2,000	2,000
1.1	4,090E+012	311381,1	4,202	-17,518
1.2	4,090E+012	169232,9	5,713	-1,755
1.3	2,302E+012	83751,041	2,579	1,475
2.0	2,302E+012	83751,041	2,579	1,475
2.1	1,066E+012	180861,4	3,347	,621
3.0	1,066E+012	180861,4	3,347	,621
3.1	3,219E+010	338598,2	4,304	1,085
4.0	3,219E+010	338598,2	4,304	1,085
4.1	2,156E+010	343521,4	5,211	1,175
5.0	2,156E+010	343521,4	5,211	1,175
5.1	2,062E+010	342437,4	5,847	1,323
6.0	2,062E+010	342437,4	5,847	1,323
6.1	2,053E+010	342318,2	6,075	1,374
7.0	2,053E+010	342318,2	6,075	1,374
7.1	2,052E+010	342262,1	6,134	1,387
8.0	2,052E+010	342262,1	6,134	1,387
8.1	2,052E+010	342247,3	6,148	1,390
9.0	2,052E+010	342247,3	6,148	1,390
9.1	2,052E+010	342243,9	6,151	1,390
10.0	2,052E+010	342243,9	6,151	1,390
10.1	2,052E+010	342243,2	6,152	1,390
11.0	2,052E+010	342243,2	6,152	1,390
11.1	2,052E+010	342243,0	6,152	1,390

Las derivadas se calculan numéricamente.

a. El número de iteraciones mayores se muestra a la izquierda del decimal, mientras que el número de iteraciones menores se encuentra a la derecha del decimal.

b. La ejecución se detuvo después de 24 evaluaciones de modelos y 11 evaluaciones de derivadas, ya que la reducción relativa entre sumas residuales sucesivas de cuadrados es, como mucho, SSCON = 1,00E-008.

CUADRO VIII.3 *(Continuación)*

En la tabla «Estimaciones de los parámetros» se muestra el valor de las constantes que en este caso serían a = 342243, b = 6,15 y c = 1,39.

Estimaciones de los parámetros

Parámetro	Estimación	Error típico	Intervalo de confianza al 95%	
			Límite inferior	Límite superior
a	342243,013	3853,134	334482,403	350003,6
b	6,152	,608	4,927	7,376
c	1,390	,137	1,114	1,667

Por lo tanto la ecuación quedaría de la siguiente forma:

$$Abundancia = \frac{342243}{1 + e^{(6{,}15 - 1{,}39 * Tiempo)}}$$

En la parte inferior de la tabla «ANOVA» observamos el Coeficiente de determinación «R cuadrado» (r^2), que es un buen indicador de la bondad del ajuste e indica la proporción de la varianza de la variable dependiente explicada por la regresión: en este caso su valor es 0,976 por lo que la regresión explica el 97,6% de la variabilidad, y tan solo el 2,4% restante es variabilidad residual no explicada por el ajuste. Un valor igual a 1 se obtendría únicamente cuando todos los puntos están en la curva, y por lo tanto el modelo explica exactamente los datos de la muestra, y en general se consideran satisfactorios porcentajes superiores a 0,90 ó 0,95.

La tabla de ANOVA contiene las sumas de cuadrados (de la regresión y residual) que permiten calcular el coeficiente r^2, así como los cuadrados medios. No se construye con ellos el cociente, estadístico que permitiría verificar la significación estadística de la regresión, ya que en general no se cumplen -en los modelos no lineales- las hipótesis necesarias para asegurar la distribución F del estadístico resultante.

ANOVA[a]

Origen	Suma de cuadrados	gl	Mean Squares
Regresión	4,07E+012	3	1,4E+012
Residual	2,05E+010	45	4,6E+008
Total sin corrección	4,09E+012	48	
Total corregido	8,42E+011	47	

Variable dependiente: Células

a. R cuadrado = 1 - (Suma de cuadrados residual) / (Suma corregida de cuadrados) = ,976.