Análisis de datos

IV. Regresión múltiple

2º de Biología

Departamento de Matemáticas

Universidad Autónoma de Madrid

2011/12

Regresión múltiple

Extensión a k variables de las ideas y técnicas de la regresión simple

- $k+1$ variables cuantitativas
- variable respuesta Y
- k variables explicativas $X_1, \ldots, X_k$
- se explica Y mediante una función lineal
  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

Ejemplos

- Supervivencia (en horas) de la larva del gusano de seda en función de la dosis de As y del peso de la larva.
- Masa de madera (en una especie de árbol) en función de la altura y del diámetro (a una determinada altura sobre el suelo).
- Tasa de respiración de un liquen en función de la concentración de Zn y la concentración de K en el agua de goteo,

Modelo

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U \qquad U \sim \mathsf{N}(0, \sigma)$$

Muestra aleatoria: $\{(y_i, x_{1i}, x_{2i}, \ldots, x_{ki}) \quad | \quad i = 1, 2, \ldots, n\}$
Matricialmente:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \ldots & X_{k1} \\ 1 & X_{12} & X_{22} & \ldots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$$

La matriz $\mathbf{X}$ se denomina *matriz del diseño*

Parámetros: $\beta_0, \beta_1, \ldots, \beta_k, \sigma$

Requisitos

- Linealidad
- Normalidad
- Homocedasticidad
- Independencia
- No colinealidad: *ninguna X_i puede ser combinación lineal de algunas otras*
- Número de datos mayor que el número de variables ($n \geq k + 2$)

Datos y estimación de parámetros

Geométricamente, la nube de puntos está ahora en un espacio de dimensión $k+1$. ¡Difícil de visualizar! Únicamente cuando $k=2$ podemos visualizarla en el espacio de 3 dimensiones.

Los datos se pueden organizar en un cuadro:

Datos	Y	X_1	X_2	$\ldots$	X_k
1	y_1	x_{11}	x_{21}	$\ldots$	x_{k1}
2	y_2	x_{12}	x_{22}	$\ldots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{1n}	x_{2n}	$\ldots$	x_{kn}

Cada fila representa un caso, cada columna una variable.

Con los valores que los datos proporcionan a la matriz del diseño $\mathbf{X}$ y a la matriz columna $\mathbf{Y}$ se estima: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

EJERCICIO: Compruébese esta fórmula en el caso $k=1$

Residuos

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki})$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2$$

Intervalos de confianza

Intervalos para las β_i:

$$\text{IC}_{1-\alpha}(\beta_i) = \left(\hat{\beta}_i \pm t_{n-k-1;\alpha/2} S_R \sqrt{q_{i+1,i+1}}\right)$$

donde $q_{i+1,i+1}$ es el elemento en la posición $i+1$ en la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$.
El factor $S_R\sqrt{q_{i+1,i+1}}$ es el error típico en la estimación $\hat{\beta}_i$.
El intervalo de confianza puede utilizarse para realizar un contraste de hipótesis sobre β_i:

$$\text{H}_0 \equiv \beta_i = 0$$

$$\mathcal{R} = \left\{ \left| \frac{\hat{\beta}_i}{S_R\sqrt{q_{i+1,i+1}}} \right| > t_{n-k-1;\frac{\alpha}{2}} \right\}$$

Contraste de la regresión

En regresión múltiple se puede realizar un contraste conjunto sobre los parámetros $\beta_1, \beta_2, \ldots, \beta_k$. Las hipótesis son:

$$\mathrm{H}_0 \equiv \beta_1 = \beta_2 = \cdots = \beta_k \quad ; \qquad \mathrm{H}_1 \equiv \exists i \in \{1, 2, \ldots, k\} : \beta_i \neq 0$$

OBSERVACIÓN: El parámetro β_0 está excluido del contraste

El contraste se realiza a través de un estadístico F obtenido mediante un análisis de la varianza

Análisis de la varianza

Descomposición de la suma de cuadrados total:

$$\text{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + \left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right) = \text{SCR} + \text{SCE}$$

OBSERVACIÓN: La suma de los productos $\sum(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ se anula.

Estadístico F:

$$F = \frac{\frac{\text{SCE}}{k}}{\frac{\text{SCR}}{n-k-1}}$$

Coeficiente de determinación:

$$R^2 = \frac{\text{SCE}}{\text{SCT}}$$

Tabla ANOVA

Variación	Suma de cuadrados	g. l.	Varianza	F
Explicada	SCE	k	$\frac{\text{SCE}}{k}$	$\frac{(n-k-1)\cdot\text{SCE}}{k\cdot\text{SCR}}$
Residual	SCR	$n-k-1$	$\frac{\text{SCR}}{(n-k-1)}$	
Total	SCT	$n-1$		

El estadístico F tiene distribución $F_{k,n-k-1}$.

Su expresión en términos de R^2 es:

$$F = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$$

Contrastes individuales sobre los parámetros

Tras aceptar la validez de la regresión múltiple realizamos contrastes individuales sobre cada una de las β_i:

$$\begin{aligned} H_0 &\equiv \quad \beta_i = 0 \\ H_1 &\equiv \quad \beta_i \neq 0 \end{aligned}$$

Utilizamos el intervalo de confianza para β_i:

$$IC_{1-\alpha}(\beta_i) = \left(\hat{\beta}_i \pm t_{n-k-1;\alpha/2} S_R \sqrt{q_{i+1,i+1}}\right)$$

RECUÉRDESE: $S_R \sqrt{q_{i+1,i+1}}$, error típico en la estimación de β_i; $q_{i+1,i+1}$ elemento de la diagonal de $(X'X)^{-1}$.

De forma análoga:

$$\mathcal{R} = \left\{ \left| \frac{\hat{\beta}_i}{S_R \sqrt{q_{i+1,i+1}}} \right| > t_{n-k-1;\alpha/2} \right\}$$

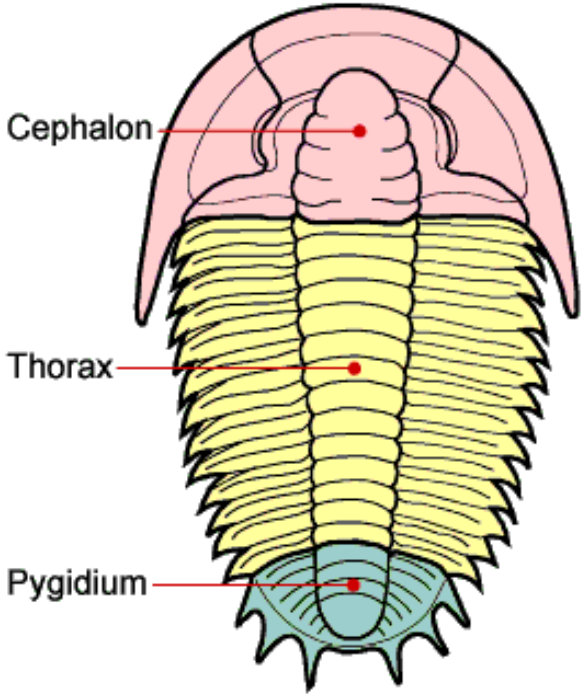
Posibles Estrategias

Contraste conjunto (F)	**Contrastes** t	**Decisión**
Modelo explicativo	Todas las X_i son explicativas	Utilizamos todas las X_i
Modelo explicativo	Algunas X_i son explicativas	Utilizamos las X_i explicativas
Modelo explicativo	Ninguna de las X_i es explicativa	Posible colinealidad, revisamos el modelo
Modelo no explicativo	Todas las X_i son explicativas	Posible colinealidad, revisamos el modelo
Modelo no explicativo	Algunas X_i son explicativas	Posible colinealidad, revisamos el modelo
Modelo no explicativo	Ninguna de las X_i es explicativa	El modelo no parece útil para explicar Y

Ejemplo 1

Estimación del tamaño de Trilobites

En la mayoría de las condiciones de preservación, es difícil encontrar ejemplares completos de Trilobites. La cabeza (cephalon) suelta es mucho más común. Por ello, es útil poder estimar el tamaño del cuerpo en función de medidas sobre la cabeza, estableciendo cuáles de ellas constituyen la mejor determinación del tamaño total.



Dibujo de Sam Gong III

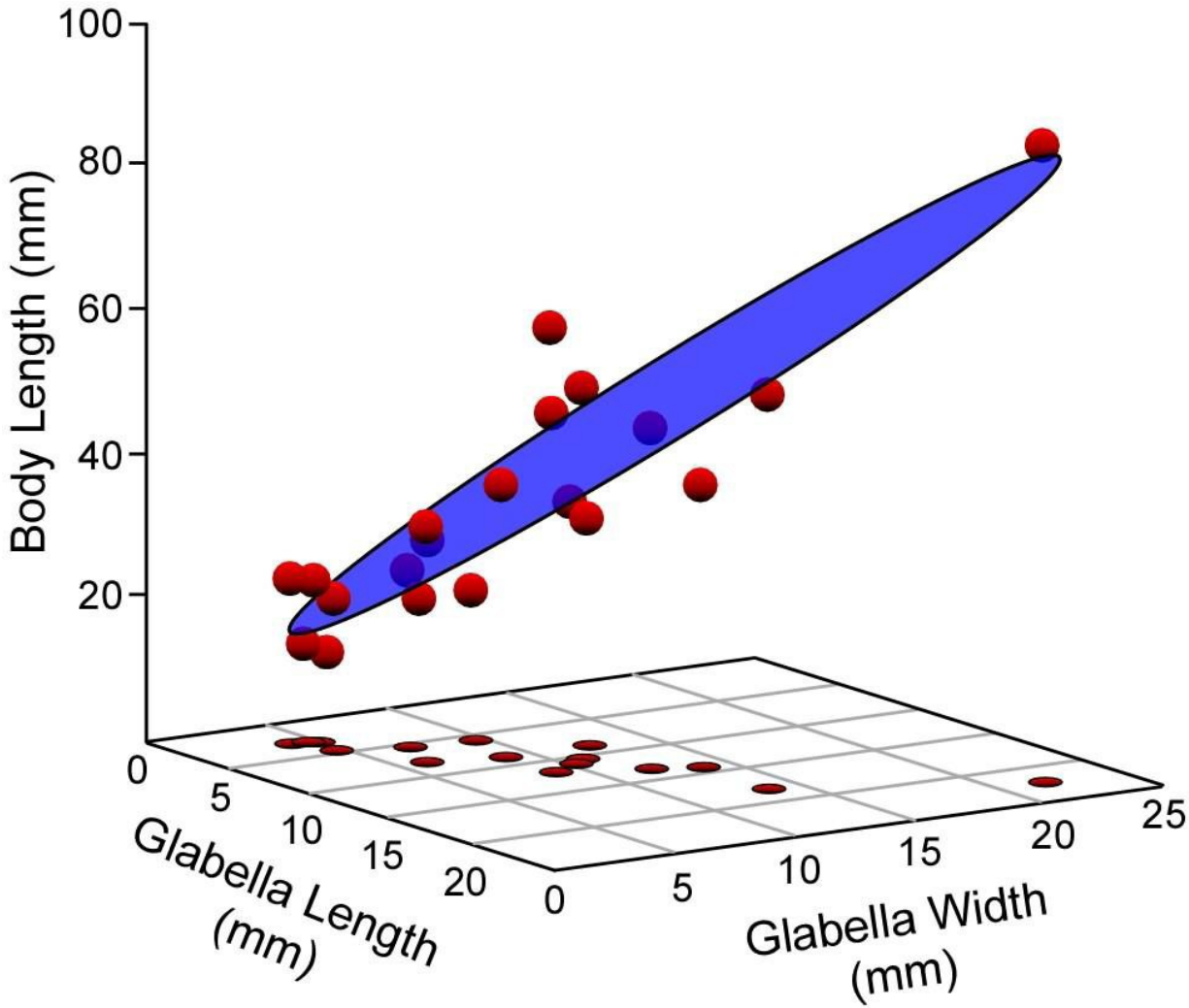
EL EJEMPLO PROCEDE DE: Norman MacLeod, Keeper of Palaeontology; The Natural History Museum, London

Datos

Table 1. Trilobite Data[1]

Genus	Body Length (mm)	Glabellar Length (mm)	Glabellar Width (mm)
Acaste	23.14	3.50	3.77
Balizoma	14.32	3.97	4.08
Calymene	51.69	10.91	10.72
Ceraurus	21.15	4.90	4.69
Cheirurus	31.74	9.33	12.11
Cybantyx	36.81	11.35	10.10
Cybeloides	25.13	6.39	6.81
Dalmanites	32.93	8.46	6.08
Delphion	21.81	6.92	9.01
Ormathops	13.88	5.03	4.34
Phacopdina	21.43	7.03	6.79
Phacops	27.23	5.30	8.19
Placopoaria	38.15	9.40	8.71
Pricyclopyge	40.11	14.98	12.98
Ptychoparia	62.17	12.25	8.71
Rhenops	55.94	19.00	13.10
Sphaerexochus	23.31	3.84	4.60
Toxochasmops	46.12	8.15	11.42
Trimerus	89.43	23.18	21.52
Zacanthoides	47.89	13.56	11.78
Mean	36.22	9.37	8.98
Std. Deviation	18.63	5.23	4.27

Gráfica


Body Length (mm)
100
80
60
40
20
0
5
10
15
20
Glabella Length (mm)
0
5
10
15
20
25
Glabella Width (mm)

Resultados

Coeficientes

	Coeficientes	Error típico	Estadístico t	p-valor	Inferior 95%	Superior 95%
Intercepción	3,9396	4,4531	0,8847	0,3887	-5,4558	13,3349
Gabella length	2,5664	0,8771	2,9259	0,0094	0,7159	4,4170
Glabella width	0,9387	1,0730	0,8749	0,3938	-1,3250	3,2025

Tabla ANOVA

	Gr. de libertad	Suma de cuadrados	cuadrados medios	F	Valor crítico de F
Regresión	2	5586,22	2793,11	40,32	0,0000004
Residuos	17	1177,70	69,28		
Total	19	6763,92			

$R^2 = 0{,}826$

Table 2. Trilobite Measurement Correlation Matrix

	y(BL)	x_1 (GL)	x_2 (GW)
y (BL)	1.000	0.895	0.859
x_1 (GL)	0.895	1.000	0.909
x_2 (GW)	0.859	0.909	1.000

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,909
Coeficiente de determinación R^2	0,826
R^2 ajustado	0,805
Error típico	8,323
Observaciones	20

Regresión simple. Variable explicativa: *Glabelar length*

Regression Statistics	
Multiple R	0.8954
R^2	0.8018
Std Error	8.5190
Adjusted R^2	0.7908
Observations	20.0000

ANOVA

	df	SS	MS	F	p-Value
Regression	1	5284.64	5285	72.82	0.0000
Residual	18	1306.31	72.57		
Total	19	6590.95			

	Coefficients	Std Error	t-Statistics	p-Value	Lower 95 %	Upper 95 %
Intercept	6.3209	3.9880	1.5850	0.1304	-2.0576	14.6994
GL	3.1900	0.3738	8.5334	0.0000	2.4046	3.9754

Regresión simple. Variable explicativa: *Glabelar width*

Regression Statistics	
Multiple R	0.8595
R^2	0.7387
Std Error	9.7815
Adjusted R^2	0.7242
Observations	20

ANOVA

	df	SS	MS	F	p-Value
Regression	1	4868.76	4869	50.89	0.0000
Residual	18	1722.19	95.68		
Total	19	6590.94			

	Coefficients	Std Error	t-Statistics	p-Value	Lower 95 %	Upper 95 %
Intercept	2.6013	5.1955	0.5007	0.6227	-8.3139	13.5165
GW	3.7455	0.5251	7.1335	0.0000	2.6424	4.8486

Ejemplo 2: Respiración de líquenes

Se estudia la tasa de respiración (en nmoles O_2 $g^{-1}min^{-1}$) del liquen *Parmelia saxatilis* en crecimiento bajo puntos de goteo con un recubrimiento galvanizado.
El agua que cae sobre el liquen contiene Zinc y Potasio que se utilizan como variables explicativas.

LOS DATOS PROCEDEN DE: Wainwright (1993), J.Biol.Educ., 27(3), 201-204.

Datos

Respiration Rate	Potassium ppm	Zinc ppm
71	388	2414
53	258	10693
55	292	11682
48	205	12560
69	449	2464
84	331	2607
21	114	16205
68	580	2005
68	622	1825

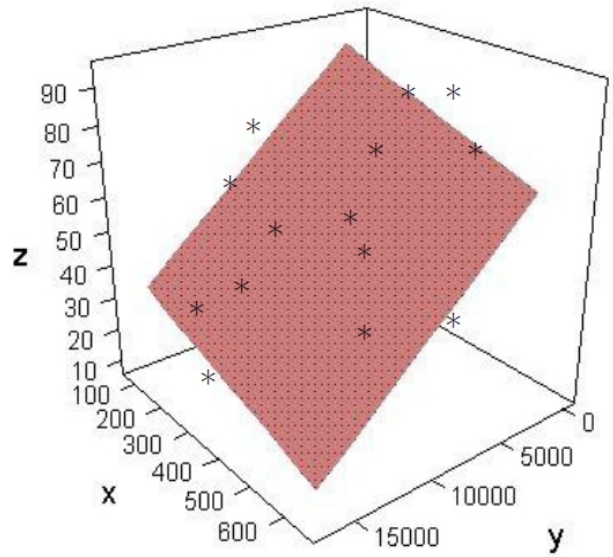


Tabla ANOVA

Source	df	SS	MS	F	p
Regression	2	2243.3	1121.6	16.80	0.003
Error	6	400.7	66.8		
Total	8	2644.0			

Coeficientes

Predictor	Coef	Stdev	t-ratio	p
Constant	101.09	18.87	5.36	0.002
K ppm	-0.04034	0.03424	-1.18	0.283
Zn ppm	-0.00387	0.001002	-3.87	0.008

Regresiones simples

ANÁLISIS DE VARIANZA (solo Zn)					
	gr. Libertad	Suma de cuadrados	cuadrados medios	F	p-valor
Regresión	1	2150,58	2150,58	30,51	0,00088423
Residuos	7	493,42	70,49		
Total	8	2644			

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,90
Coeficiente de determinación R^2	0,81
R^2 ajustado	0,79
Error típico	8,40
Observaciones	9

ANÁLISIS DE VARIANZA (sólo K)					
	gr. Libertad	Suma de cuadrados	cuadrados medios	F	p-valor
Regresión	1	1244,51	1244,51	6,22	0,04
Residuos	7	1399,49	199,93		
Total	8	2644			

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,69
Coeficiente de determinación R^2	0,47
R^2 ajustado	0,40
Error típico	14,14
Observaciones	9

Interacción

Regression Statistics	
Multiple R	0.9812
R^2	0.9628
Standard Error	4.4372
Adjusted R^2	0.9404
Observations	9

ANOVA	df	SS	MS	F	p-Value
Regression	3	2545.55	848.52	43.10	0.0005
Residual	5	98.45	19.69		
Total	8	2644.00			

	Coeffs	Std Error	t-Statistics	p-Value
Intercept	106.07	10.32	10.28	0.0002
K ppm	-0.0678	0.0199	-3.4129	0.0190
Zn ppm	-0.0060	0.0008	-7.8124	0.0006
K$\times$Zn	1.1184	0.2854	3.9181	0.0112

Ejemplo 3: Supervivencia de la larva del gusano de seda

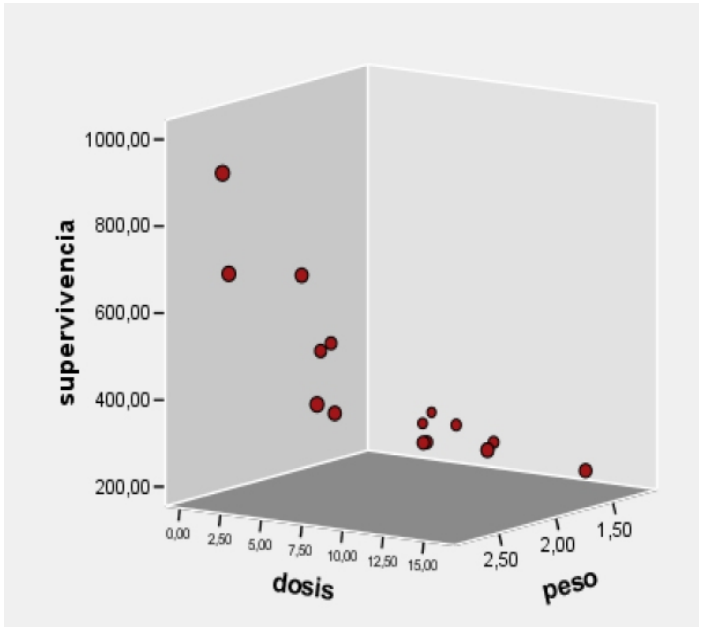
En un experimento sobre el efecto tóxico del arsénico sobre las larvas del gusano de seda[1], se inyectaron distintas dosis del compuesto químico a 15 larvas de distintos pesos, midiéndose posteriormente su supervivencia.

[1] Un estudio recurrente tras F. L. CAMPBELL, 1926

Datos

Supervivencia	dosis	peso
685,49	1,41	2,66
924,70	1,64	2,75
486,41	3,07	2,00
477,53	3,23	2,11
671,43	3,72	2,35
276,69	3,92	1,24
263,63	4,37	1,38
399,94	6,04	2,55
359,75	5,48	2,31
276,06	6,79	1,43
263,03	7,33	1,77
274,79	8,02	1,90
242,66	8,75	1,38
283,14	12,30	1,95
224,39	15,63	1,56

Datos



A la vista de la forma de la nube de puntos transformamos los datos. Tras varios intentos las transformaciones
$Y = \log_{10}(supervivencia);$
$X_1 = \log_{10}(dosis); \quad X_2 = \log_{10}(peso);$
dan un resultado aceptable.

Datos

Y	X_1	X_2
2,84	,15	,43
2,97	,21	,44
2,69	,49	,30
2,68	,51	,33
2,83	,57	,37
2,44	,59	,09
2,42	,64	,14
2,60	,78	,41
2,56	,74	,36
2,44	,83	,16
2,42	,87	,25
2,44	,90	,28
2,39	,94	,14
2,45	1,09	,29
2,35	1,19	,19

Resultados

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresion	,464	2	,232	59,178	,000[a]
	Residual	,047	12	,004		
	Total	,511	14			

a. Variables predictoras: (Constante), Log10 (peso), Log10 (dosis)

b. Variable dependiente: Log10 (supervivencia)

		Estadísticos				
		Coeficientes no estandarizados		Coeficientes estandarizados		
Modelo	Variables	B	Error tip.	Beta	t	Sig.
1	(Constante)	2,589	,084		30,966	,000
	Log10 (dosis)	-,378	,066	-,580	-5,702	,000
	Log10 (peso)	,875	,172	,516	5,073	,000

a. Variable dependiente: Log10 (supervivencia)

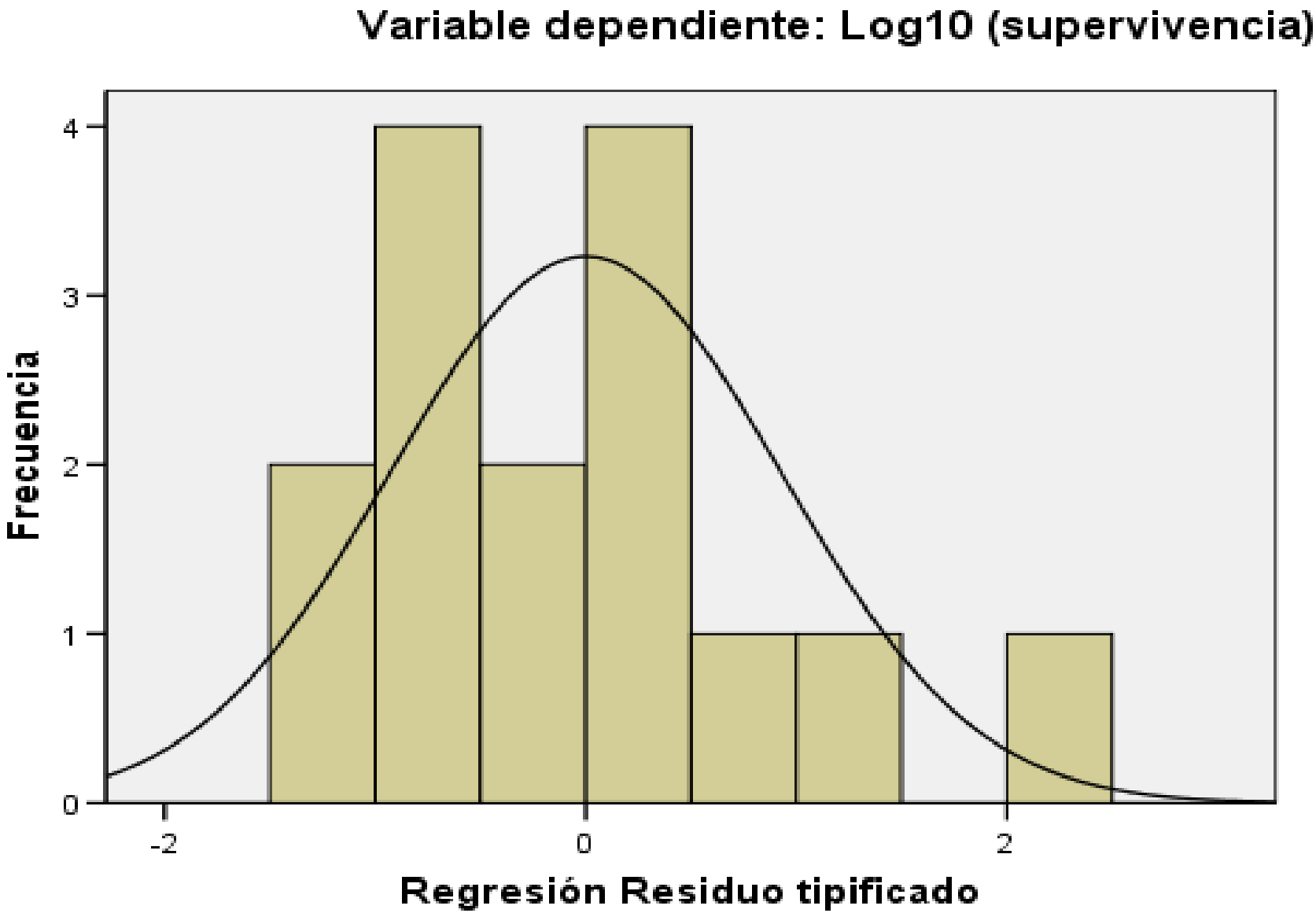
Ecuación de la regresión

$$Y = 2{,}59 - 0{,}378X_1 + 0{,}875X_2; \qquad R^2 = 0{,}91$$

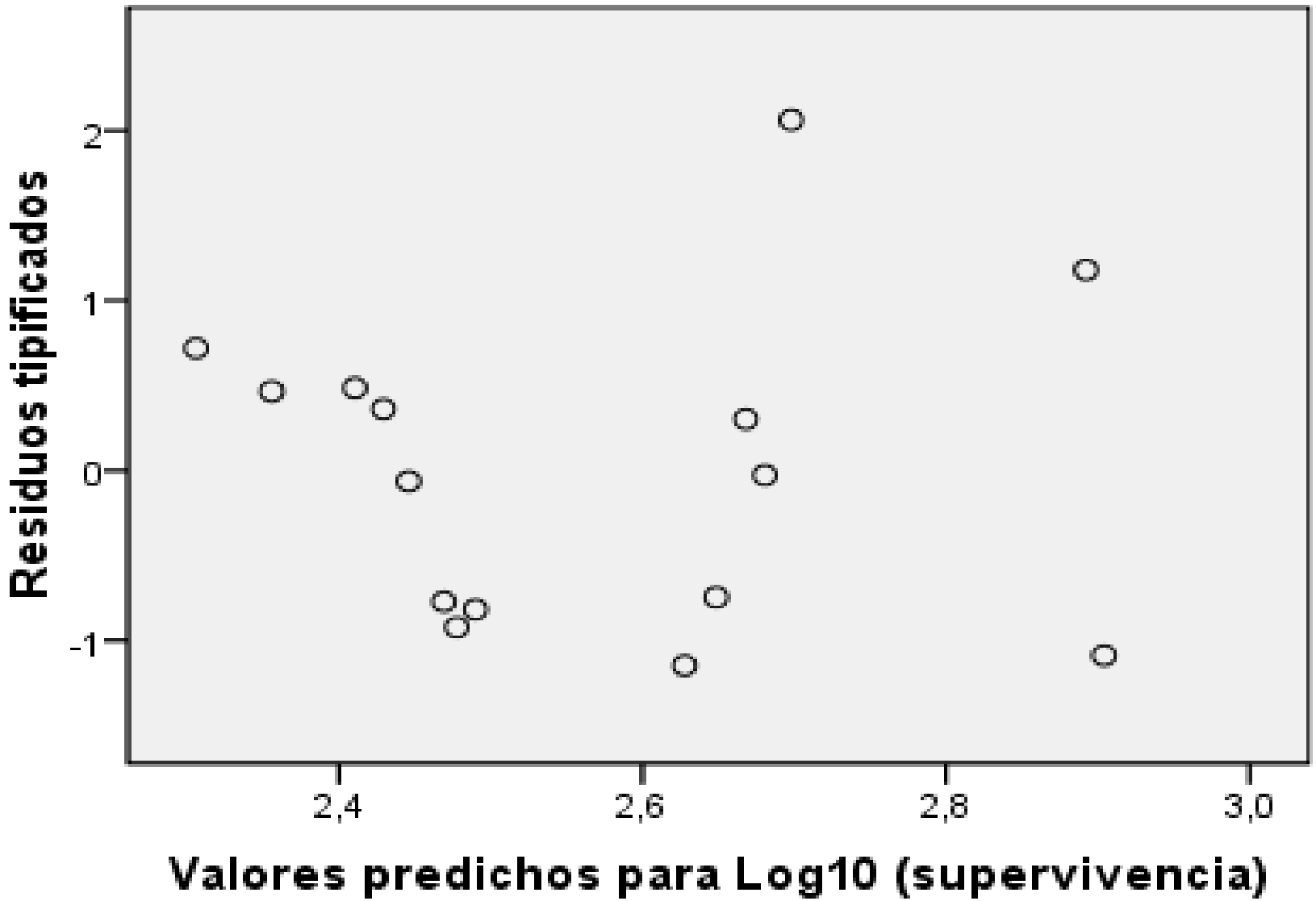
o bien

$$supervivencia = 10^{2{,}59}(dosis)^{-0{,}378}(peso)^{0{,}875}$$

Histograma de residuos



Dispersión de residuos



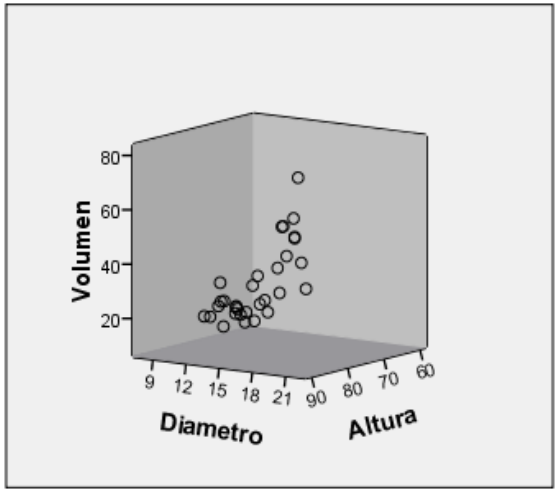
Ejemplo 4. Volumen del cerezo negro

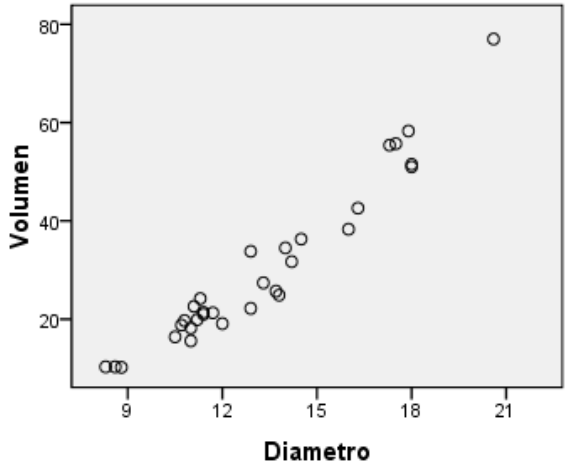
Los siguientes datos proceden de un análisis realizado sobre el cerezo negro en Allegheny National Forest, Pennsylvania. Las variables, medidas en 31 cerezos, son:

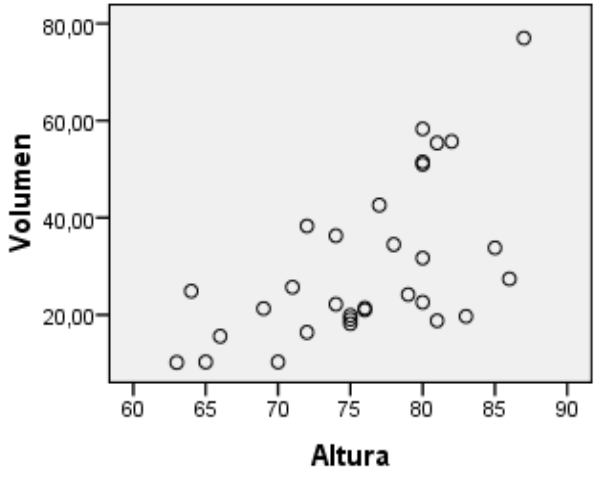
- volumen (en pies cúbicos)
- altura (en pies)
- diámetro (en pulgadas, a 54 pulgadas sobre la base)

Se trata de estimar el volumen de un árbol (y por tanto su cantidad de madera) dados su altura y su diámetro.

Datos: representación gráfica







Descriptivos

Estadísticos descriptivos

	Media	Desviación típ.	N
Volumen	30,1710	16,43785	31
Diametro	13,2484	3,13814	31
Altura	76,00	6,372	31

Resumen del modelo[b]

Modelo	R	R cuadrado
1	,974[a]	,948

a. Variables predictoras: (Constante), Altura, Diametro

b. Variable dependiente: Volumen

Correlaciones			
	Diam	*Altura*	*Volumen*
Diámetro	1		
Altura	0,519	1	
Volumen	0,967	0,598	1

Varianzas y covarianzas			
	Diam	*Altura*	*Volumen*
Diámetro	7,986		
Altura	7,598	36,432	
Volumen	38,030	44,917	194,668

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	7684,163	2	3842,081	254,972	,000[a]
	Residual	421,921	28	15,069		
	Total	8106,084	30			

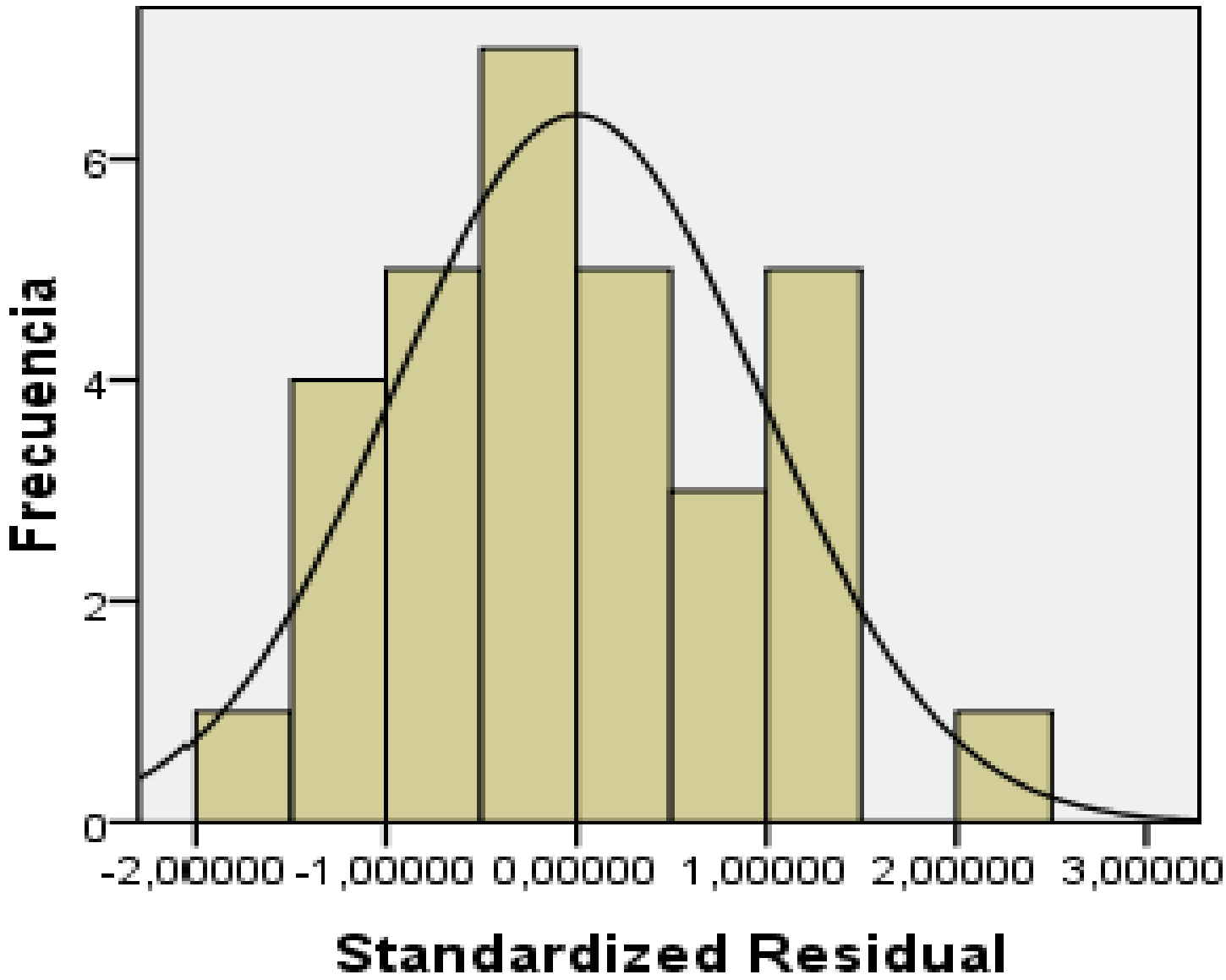
a. Variables predictoras: (Constante), Altura, Diametro

Coeficientes[a]

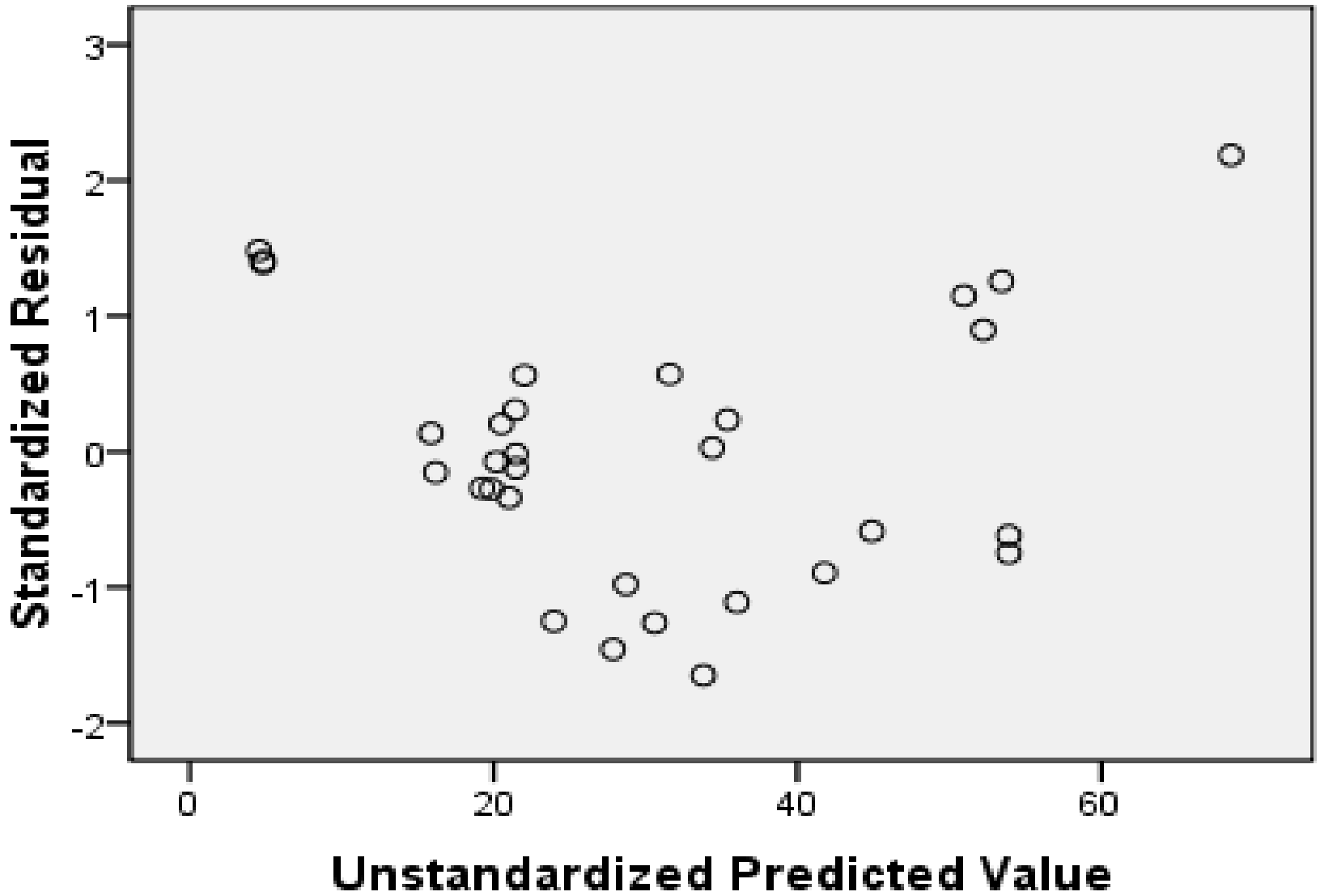
Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	-57,988	8,638		-6,713	,000		
	Diametro	4,708	,264	,899	17,816	,000	,730	1,369
	Altura	,339	,130	,132	2,607	,014	,730	1,369

a. Variable dependiente: Volumen

Histograma de residuos



Dispersión de residuos



Incorporación de una variable dicotómica

- Además de las k variables continuas, tenemos una variable cualitativa con dos valores (codificados como 0 y 1)
- Esta nueva variable puede incorporarse a la regresión.
- Se obtiene un coeficiente adicional β_{k+1}.
- El análisis de la regresión nos indica si el efecto de esta variable es o no significativo.
- El coeficiente β_{k+1} representa lo que se añade a la ecuación lineal $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ cuando la nueva variable toma el valor 1.

Ejemplo: «Peso» en función de la «Estatura» con la variable «sexo» como dicotómica

Estudiamos las rectas de regresión del «Peso» sobre la «Estatura» primero con todos los datos, después con hombres y mujeres por separado y finalmente con todos los datos y utilizando como segunda variable explicativa el sexo, codificada con "0"para hombres y "1"para mujeres.

Con todos los datos juntos

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4823,302	1	4823,302	89,306	,000[a]
	Residual	3996,632	74	54,009		
	Total	8819,934	75			

a. Variables predictoras: (Constante), estatura
b. Variable dependiente: peso

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-103,322	17,396		-5,939	,000
	estatura	,975	,103	,740	9,450	,000

a. Variable dependiente: peso

Datos de hombres

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	641,300	1	641,300	7,971	,009[a]
	Residual	1930,854	24	80,452		
	Total	2572,154	25			

a. Variables predictoras: (Constante), estatura
b. Variable dependiente: peso

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-72,765	51,087		-1,424	,167
	estatura	,816	,289	,499	2,823	,009

a. Variable dependiente: peso

Datos de mujeres

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	397,854	1	397,854	12,814	,001[a]
	Residual	1490,326	48	31,048		
	Total	1888,180	49			

a. Variables predictoras: (Constante), estatura
b. Variable dependiente: peso

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-30,225	23,938		-1,263	,213
	estatura	,521	,146	,459	3,580	,001

a. Variable dependiente: peso

Con sexo como variable dicotómica

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	5348,415	2	2674,207	56,234	,000[a]
	Residual	3471,519	73	47,555		
	Total	8819,934	75			

a. Variables predictoras: (Constante), sexo, estatura
b. Variable dependiente: peso

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-41,405	24,772		-1,671	,099
	estatura	,638	,140	,484	4,560	,000
	sexo	-8,016	2,412	-,353	-3,323	,001

a. Variable dependiente: peso

Extensión a un modelo de regresión logística

A partir de una colección de observaciones $(y_i, x_{i1}, \ldots, x_{ik})$ de las variables $Y, X_1, \ldots, X_k$, donde la variable y es dicotómica (con valores 0 y 1) y las variables $X_1, \ldots, X_k$ son continuas, buscamos una forma de decidir a partir de valores observados $(x_1, \ldots, x_k)$ si el valor de la variable y debe ser 0 o 1.

Ejemplo introductorio

El 4 de julio de 1999 una tormenta con vientos que excedían los 140 km/h azotó el NE de Minnesota causando graves daños en los bosques de un parque natural. Analizados los efectos de la tormenta sobre más de 3.600 árboles del parque se midieron las siguientes variables:

- Diámetro en cm (variable D)
- Medida de la fuerza de la tormenta relacionada con el porcentaje inerte de área basal[2] (variable S)
- Registro de si cada árbol había muerto ($Y = 1$) o si había sobrevivido ($Y = 0$)
- Especie a la que pertenecía cada árbol (variable SSP).

Tras un primer análisis descriptivo, parece que el diámetro D y la fuerza de la tormenta S pueden ser útiles para estimar la probabilidad de supervivencia de un árbol (Y).

[2] Área de un terreno ocupada por la sección de los troncos de los árboles en la base.

El modelo de regresión logística que veremos se utiliza para investigar la relación entre una variable respuesta cualitativa que toma dos posibles valores (en el ejemplo, la variable Y) y un conjunto de variables regresoras continuas (en el ejemplo, las variables D, S).

Este modelo nos permitirá:

- describir la probabilidad de que un árbol sobreviva o no como función del resto de las variables explicativas,
- determinar si estas variables modifican significativamente dicha probabilidad (influyen en la Y)
- estimar en función de las variables regresoras la probabilidad de que un árbol sobreviva o no.

Descripción del modelo

Tenemos una variable Y con dos valores ($0 =$ fracaso y $1 =$ éxito) [Por ejemplo, la variable que describe si un árbol sobrevive o no.] Esta variable depende de los valores de otras $x_1, \ldots, x_k$ (continuas) [Por ejemplo las variables D y S.] Es decir: La probabilidad de éxito dependerá de los valores de las x's La fórmula para esta probabilidad viene dada por la función logística:

$$P(Y = 1|_{X_1=x_1,\ldots,X_k=x_k}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}$$

Muestra aleatoria

Realizaremos n observaciones independientes. Cada observación $(y_i, x_{i1}, \ldots, x_{ik})$ estará formada por el valor de la variable respuesta Y_i (que puede ser cero o uno) y los valores de cada una de las variables regresoras. Supondremos que: las variables $Y_1, \ldots, Y_n$ son independientes y que para cada $i = 1, 2, \ldots, n$

$$p_i = P(Y = 1|_{\mathbf{X}=\mathbf{x}_i}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik})}}$$

donde $\mathbf{x}_i$ representa $(x_{i1}, \ldots, x_{ik})$

Ejemplo: datos

Supervivencia:

$Y_i = 1$ el árbol i no sobrevivió a la tormenta

$Y_i = 0$ el árbol i sobrevivió a la tormenta

$n = 3\,666$

Diametro (D)	Fuerza (S)	Supervivencia (Y)	Especie (SPP)
9,00	0,024	0	BF
7,00	0,028	0	BA
7,00	0,102	0	BS
13,00	0,102	0	C
15,00	0,210	0	C
9,00	0,210	0	PB
20,00	0,306	0	C
16,00	0,307	0	BS
13,00	0,426	1	JP
7,00	0,429	0	BF
18,00	0,509	1	PB
37,00	0,511	0	A
16,00	0,626	1	JP
15,00	0,628	1	BS
9,00	0,716	1	BF
15,00	0,717	1	BS
17,00	0,847	1	BS
14,00	0,847	0	RM
30,00	0,974	1	RP
8,00	0,983	1	BF
etc.	etc.	etc.	etc.

Estimación de parámetros

La estimación de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ se realiza por métodos numéricos.
Una vez estimados estos parámetros se estiman las probabilidades:

$$\hat{p}_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}} \qquad i = 1, 2, \ldots, n$$

Análisis con SPSS

`Analizar ↪ Regresión ↪ Logística binaria...`

Como variable dependiente elegimos la variable respuesta; en la ventana «Covariables» situamos las variables D y S.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 [a]	D	,097	,005	346,022	1	,000	1,102
	S	4,424	,189	545,122	1	,000	83,412
	Constante	-3,543	,127	774,463	1	,000	,029

a. Variable(s) introducida(s) en el paso 1: D, S.

La columna B da las estimaciones de los parámetros β_i; la columna E.T. da los errores típicos de estas estimaciones.
Para el árbol i calculamos:

$$\hat{p}_i = \frac{1}{1 + e^{-(-3{,}543 + 0{,}097 D_i + 4{,}424 S_i)}}$$

La opción «Probabilidades» del botón «Guardar...» del cuadro de diálogo permite calcular las $\hat{p}_i$ de todos los árboles de la muestra.

Interpretación de los contrastes de hipótesis

En la tabla «Variables de la ecuación» que muestra SPSS aparece una columna encabezada «Wald».
Esta columna recoge el valor del estadístico de Wald y se calcula como el cuadrado de la estimación de la β dividida por su error típico.

$$\text{W} = \left(\frac{\hat{\beta}_i}{\text{error típico en } \hat{\beta}_i} \right)^2$$

Este estadístico tiene distribución χ^2 con 1 grado de libertad.
Su p-valor se da en la columna «Sig.»
Este p-valor se utiliza para decidir el contraste que tiene por hipótesis nula:

$$\text{H}_0 \equiv \beta_i = 0$$

Predicciones y clasificación

Para un valor no observado $(x_{01}, x_{02}, \ldots, x_{0k})$ se estima

$$\hat{p}_0 = P(Y = 1|_{\mathbf{X}=\mathbf{x}_0}) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k})}}$$

y se clasifica como $\hat{Y}_0 = 1$ cuando $\hat{p}_0 > \frac{1}{2}$; es decir:

$$\hat{Y}_0 = 1 \equiv \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k} > 0$$

En el ejemplo esto ocurre si

$$0{,}097 \cdot D_0 + 4{,}424 \cdot S_0 > 3{,}543$$

Ejercicio

En el ejemplo, estimar la probabilidad de que no sobreviva un árbol cuyo diámetro es de 30 cm y está situado en una zona en la que la fuerza de la tormenta viene dada por $S = 0{,}8$.

Ejemplo: *Iris sp.*

Se dispone de medidas en cm de la longitud y anchura del pétalo y de la longitud y anchura del sépalo de 100 lirios correspondientes a dos especies diferentes: *Iris versicolor* ($y = 0$) e *Iris virginica* ($y = 1$). Se ha ajustado un modelo de regresión logística a los datos con el fin de estudiar la probabilidad de que un lirio pertenezca a cada una de las dos especies en función de las cuatro medidas.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	LSepalo	-2,465	2,394	1,060	1	,303	,085
	ASepalo	-6,681	4,480	2,224	1	,136	,001
	LPetalo	9,429	4,737	3,962	1	,047	12448,870
	APetalo	18,286	9,743	3,523	1	,061	8,741E7
	Constante	-42,638	25,708	2,751	1	,097	,000

a. Variable(s) introducida(s) en el paso 1: LSepalo, ASepalo, LPetalo, APetalo.