Las técnicas de Regresión lineal multiple parten de k+1 variables cuantitativas:

La variable respuesta (y)
Las variables explicativas (x_1 ,..., x_k)

Y tratan de explicar la y mediante una función lineal de las x_1 ,..., x_k representada por:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Debemos extender a k variables las ideas y técnicas de la regresión lineal simple

Modelo

$$Y_{(x1,...xk)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + U \text{ con } U \Longrightarrow N(0,\sigma)$$

Muestra Aleatoria

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i, \text{ para } i = 1, \cdots, n.$$

$$Y_i \sim N(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \sigma^2) \text{ independientes}, i = 1, \cdots, n$$

$$u_i \sim N(0, \sigma^2) \text{ independientes}, i = 1, \cdots, n$$

En notación matricial:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$Y = X\beta + U, \qquad \textbf{X = matriz del diseño}$$

Cuatro hipótesis comunes con la regresión lineal simple

- Normalidad
- Homocedasticidad
- Linealidad
- Independencia

Y dos requisitos adicionales

- **$n > k+1$**

  El modelo depende de k+2 parámetros. Para que la regresión tenga sentido debemos tener un número suficiente de datos (evidentemente, en la regresión lineal simple, también necesitamos más de 2 datos para que tenga sentido ajustar una recta)

- **Ninguna de las X_i es combinación lineal de las otras (multicolinealidad)**

  Si alguna de las X_i es combinación lineal exacta de algunas de las otras X_j, el modelo puede simplificarse con menos variables explicativas. También hay que tener cuidado si alguna de las X_i está fuertemente correlacionada con otras.

Datos y estimación de los parámetros

Datos	Y	X_1	$\cdots$	X_k
1	y_1	x_{11}	$\cdots$	x_{k1}
2	y_2	x_{12}	$\cdots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
n	y_n	x_{1n}	$\cdots$	x_{kn}

Geométricamente, la nube de puntos ahora está en un espacio de dimensión k+1
¡Difícil de visualizar para k>2!

X es la matriz del diseño, ahora con los datos

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

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

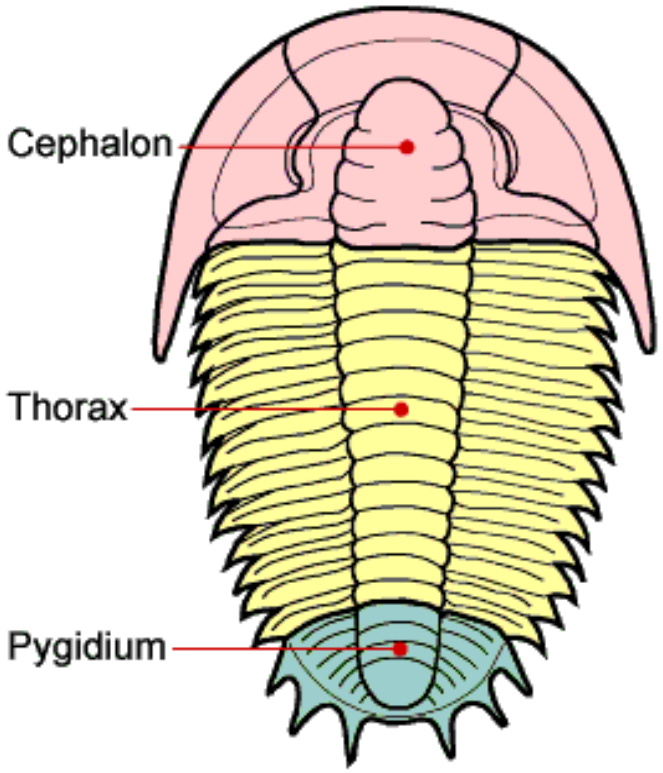
$$\hat{\sigma}^2 = S_R^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2.$$

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.

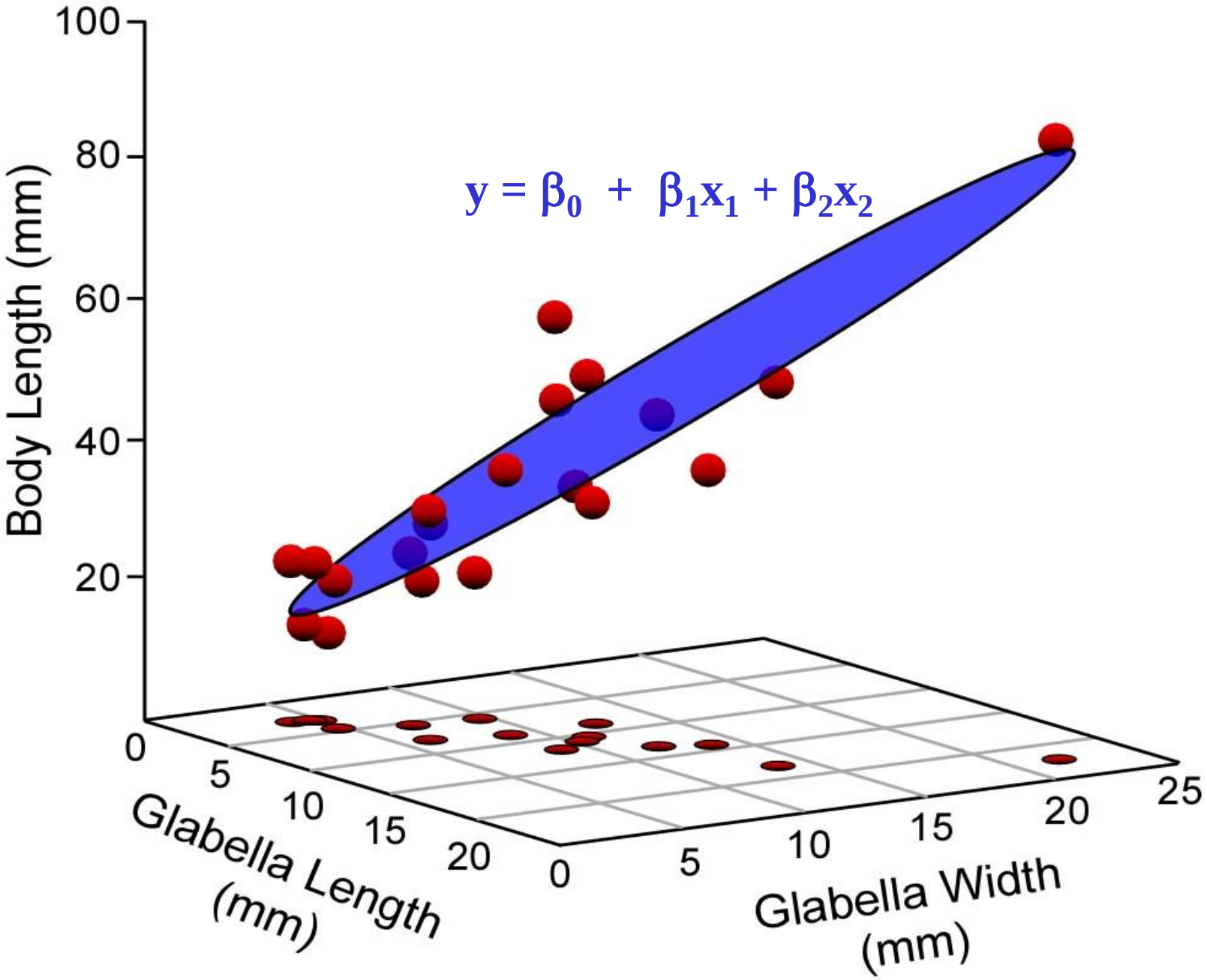
El siguiente ejemplo está tomado de:
Norman MacLeod
Keeper of Palaeontology,
The Natural History Museum, London



Dibujo de Sam Gon III

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


$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
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)

Intervalos de confianza

Para β_i, con $i = 0, 1, \cdots, k$:

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

Error típico de la estimación de $\hat{\beta}_i$

siendo $q_{i+1,i+1}$ los elementos de la diagonal principal de $(X'X)^{-1}$.

Contrastes de hipótesis

$$H_0 : \beta_i = 0 \ (X_i \text{ no influye sobre } Y)$$
$$H_1 : \beta_i \neq 0 \ (X_i \text{ influye sobre } Y)$$

Rechazaremos H_0 , al nivel α, si el cero no está en el intervalo de confianza 1- α para β_i.

Lo que es equivalente al contraste de la t de Student para cada parámetro β_i.

Ejemplo 1
Estimación del tamaño de Trilobites

	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

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

Análisis de la Varianza

$H_0 : \beta_1 = \cdots = \beta_k = 0$ (el modelo no es explicativo)

H_1 : al menos un $\beta_i \neq 0$ (el modelo es explicativo)

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

$$= \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SCR}$$

SCE (suma de cuadrados explicada)

SCR (suma de cuadrados residual)

Coeficiente de determinación $R^2 = \frac{SCE}{SCT}$.

$SCT = nv_y = (n-1)\, s_y^2$; $SCE = nv_y R^2$

Tabla Anova

Variación	Suma de cuadrados	g. l.	Media cuadrática	F
Explicada	$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	SCE/k	$\frac{SCE/k}{SCR/(n-k-1)}$
Residual	$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k-1$	$SCR/(n-k-1)$	
Total	$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

Rechazaremos H_0 , al nivel α, si :

$$F \geq F_{k,n-k-1;\alpha}$$

Relación entre F y R^2

$$F = \frac{SCE/k}{SCR/(n-k-1)} = \frac{SCE}{SCT}\frac{SCT}{SCR}\frac{n-k-1}{k} = R^2\frac{1}{\frac{SCT-SCE}{SCT}}\frac{n-k-1}{k} = \frac{R^2}{1-R^2}\frac{n-k-1}{k}.$$

Ejemplo 1
Estimación del tamaño de Trilobites

	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			

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

Resumen de los contrastes

Contraste conjunto (F)	**Contrastes individuales (t)**	**Conclusión**
Modelo explicativo	Todas las X_i son explicativas	Nos quedamos con todas las X_i
Modelo explicativo	Algunas X_i son explicativas	Nos quedamos con las X_i explicativas
Modelo explicativo	Ninguna X_i es explicativa	Posible multicolinealidad (revisar el modelo)
Modelo no explicativo	Todas las X_i son explicativas	Posible multicolinealidad (revisar el modelo)
Modelo no explicativo	Algunas X_i son explicativas	Posible multicolinealidad (revisar el modelo)
Modelo no explicativo	Ninguna X_i es explicativa	El modelo no es útil para explicar Y

Ejemplo 2
Respiración de líquenes

Se estudia la tasa de respiración (en nmoles oxígeno 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 corresponden a:
Wainwright (1993, J.Biol.Educ., 27(3), 201- 204).

Datos

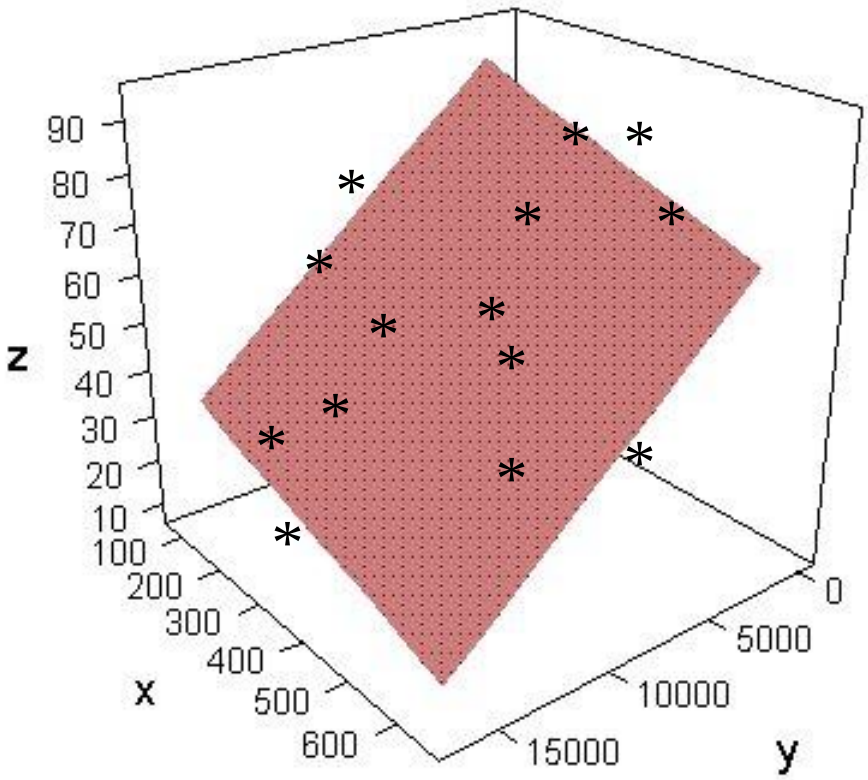
Respiration Rate	Potassium ppm	Zinc ppm
71	388	2414
53	258	10693
55	292	11682
48	205	12560
69	449	2464
84	331	2607
21	114	16205
68	580	2005
68	622	1825

Variable	N	MEAN	MEDIAN	STDEV
RespRate	9	59.67	68.00	18.8
K ppm	9	359.9	331.0	168.1
Zn ppm	9	6939	2607	5742

Plano de regresión

Tasa de respiración = $\beta_0 + \beta_1$Potasio + β_2Zinc

Datos *



Análisis de la varianza (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			

Regresión de la tasa de respiración (RespRate) sobre el Potasio (K) y el Zinc (Zn). La ecuación de regresión estimada es:

RespRate = 101 - 0.0403 K - 0.00388 Zn

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

ANÁLISIS DE VARIANZA (sólo 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

Estimación de la respuesta media de Y para los valores x_{10} ,…, x_{k0} de las variables explicativas

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots + \hat{\beta}_k x_{k0}.$$

$$IC_{1-\alpha} = \left(\hat{y}_0 \mp t_{n-k-1;\alpha/2} S_R \sqrt{\begin{pmatrix} 1 & x_{10} & \cdots & x_{k0} \end{pmatrix} (X'X)^{-1} \begin{pmatrix} 1 \\ x_{10} \\ \cdots \\ x_{k0} \end{pmatrix}} \right)$$

Error típico

Predicción de un nuevo valor de Y dados los valores x_{10} ,…, x_{k0} de las variables explicativas

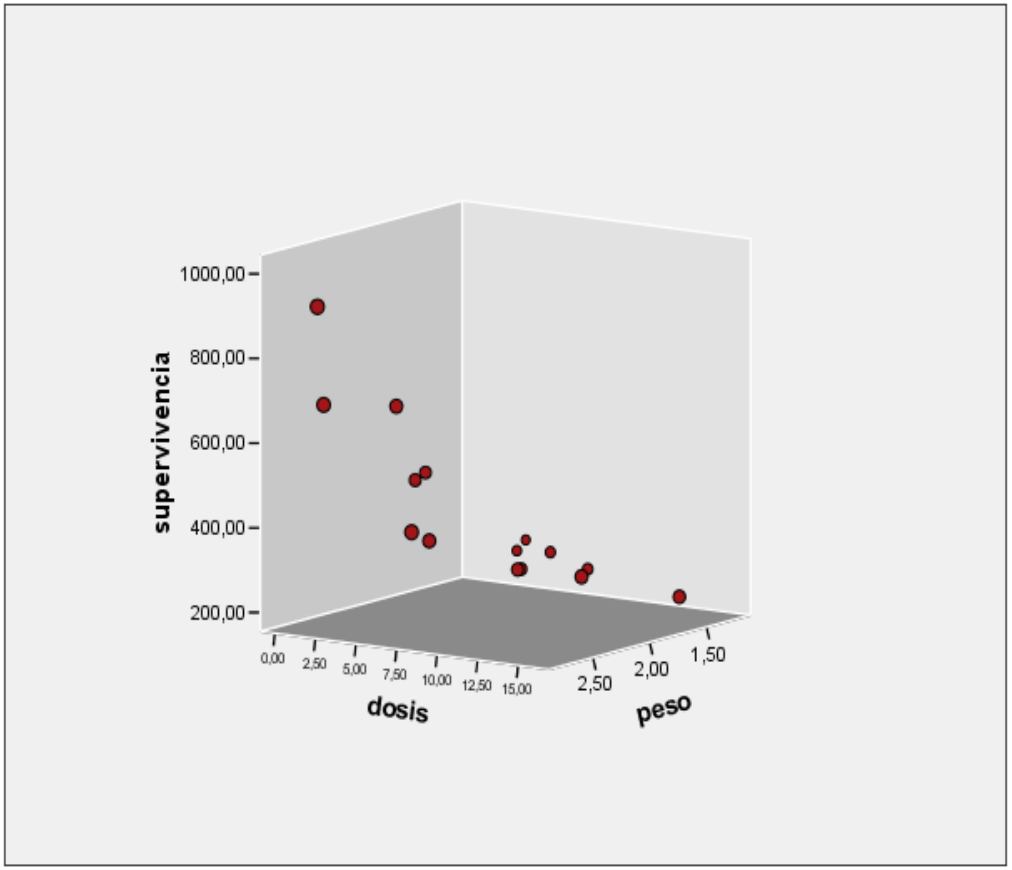
$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \cdots + \hat{\beta}_k x_{k0}.$$

$$IP_{1-\alpha} = \left(\hat{y}_0 \mp t_{n-k-1;\alpha/2} S_R \sqrt{1 + \begin{pmatrix} 1 & x_{10} & \cdots & x_{k0} \end{pmatrix} (X'X)^{-1} \begin{pmatrix} 1 \\ x_{10} \\ \cdots \\ x_{k0} \end{pmatrix}} \right)$$

Error típico

Ejemplo 3

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



Se decidió realizar una regresión lineal múltiple entre las variables:

$Y = Log_{10}$ (supervivencia)

$X_1 = Log_{10}$ (dosis)

$X_2 = Log_{10}$ (peso)

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

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

Datos transformados

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,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)

Coeficientes[a]

		Estadísticos				
		Coeficientes no estandarizados		Coeficientes estandarizados		
Modelo	Variables	B	Error típ.	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

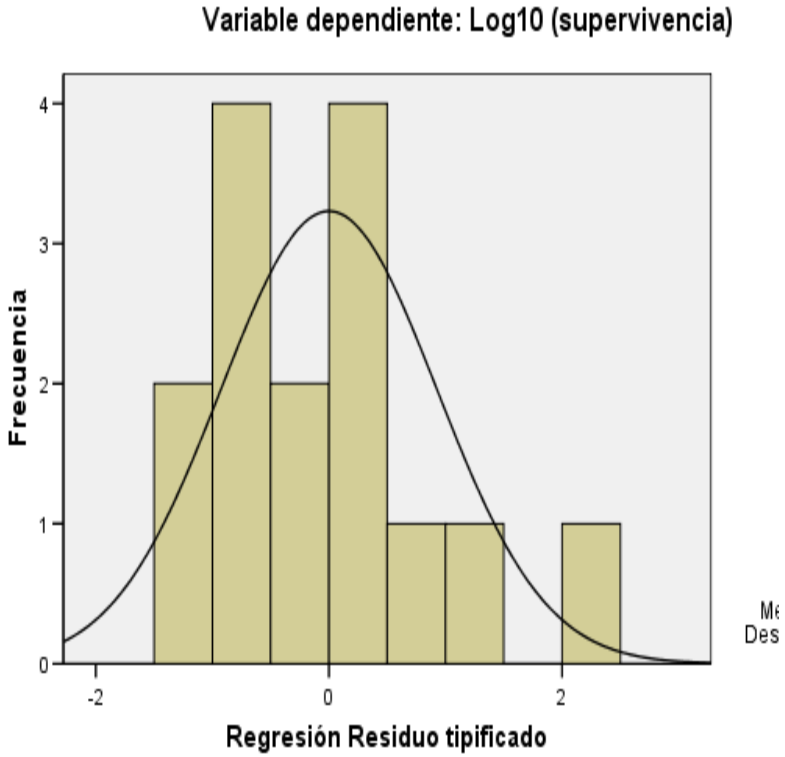
a. Variable dependiente: Log10 (supervivencia)

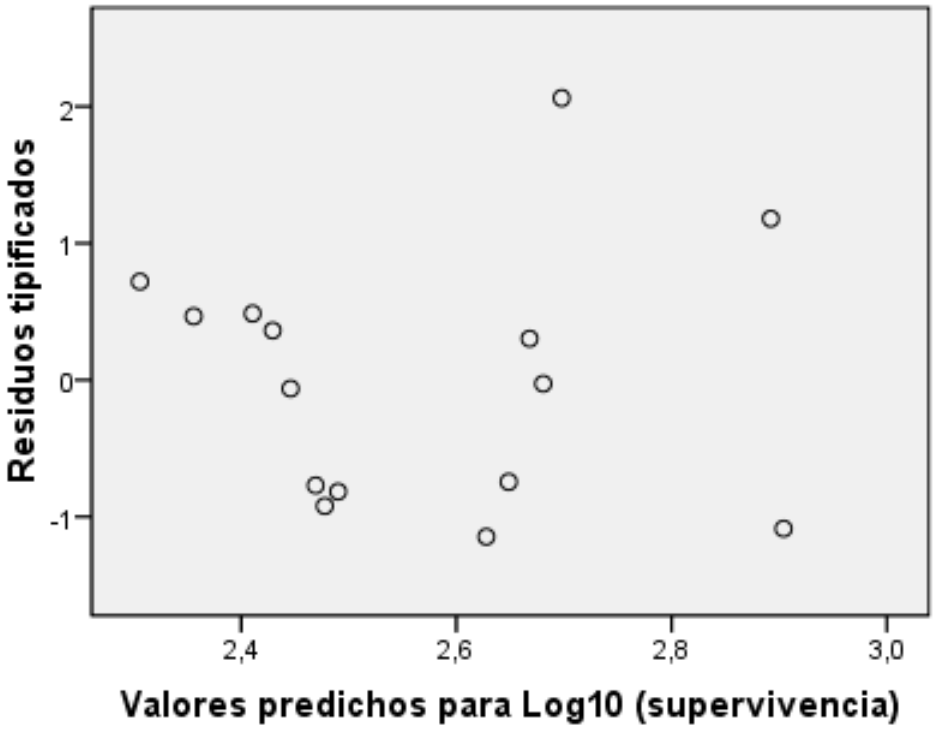
$$R^2(X_1 \text{ and } X_2) = 0.4637/0.5113 = 0.9069$$

$$R^2(X_1 \text{ alone}) = 0.3633/0.5113 = 0.7105$$

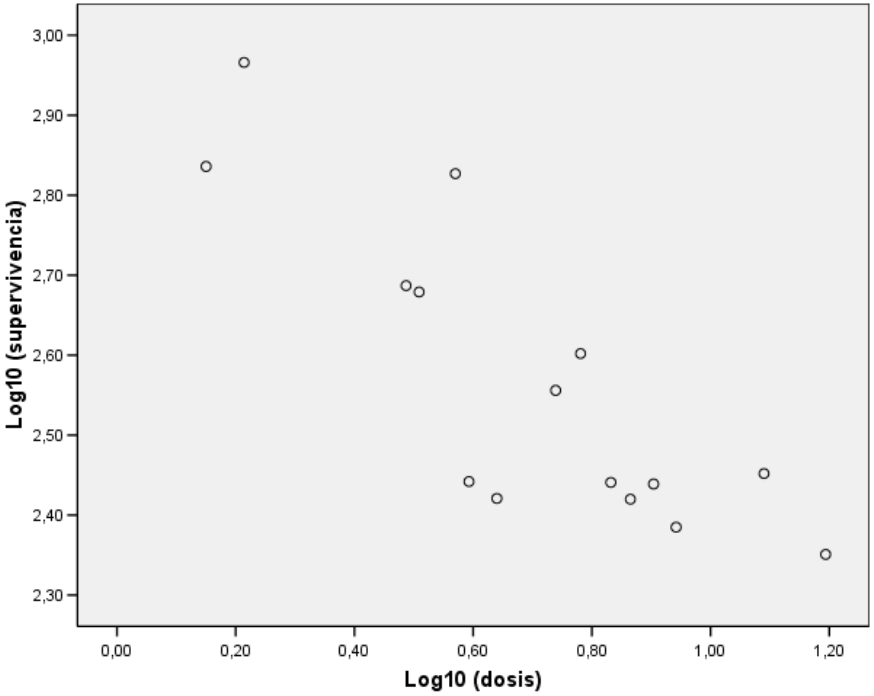
$$R^2(X_2 \text{ alone}) = 0.3332/0.5113 = 0.6517$$

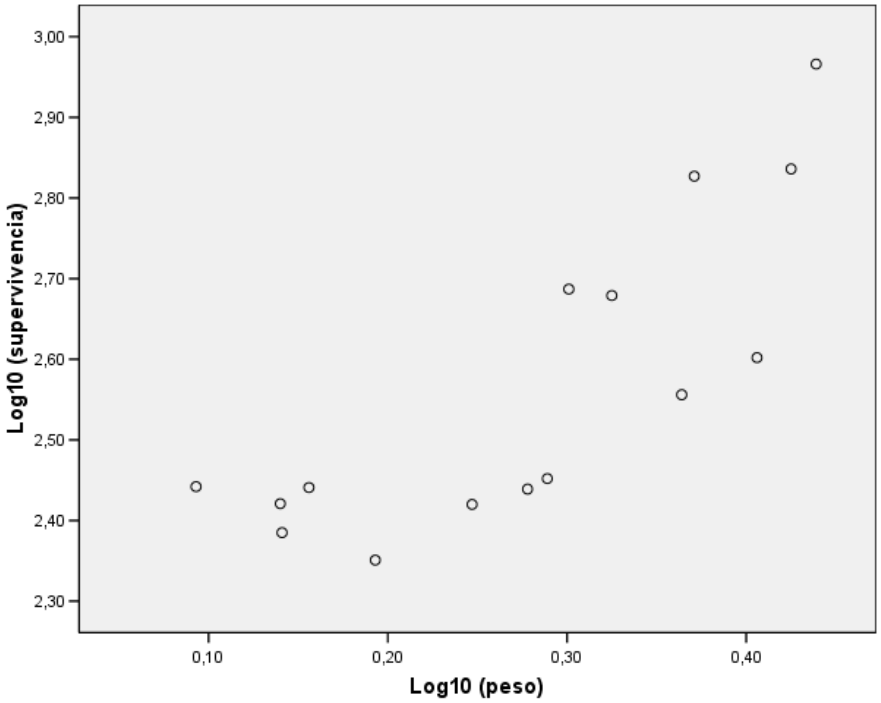
Histograma


Variable dependiente: Log10 (supervivencia)
Frecuencia
4
3
2
1
0
-2
0
2
Regresión Residuo tipificado
Me
Des


Residuos tipificados
2
1
0
-1
2,4
2,6
2,8
3,0
Valores predichos para Log10 (supervivencia)

Gráficos de regresión simple





Regresión simple: sólo la dosis

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	2,952	,074		40,136	,000	2,793	3,111
	Log10 (dosis)	-,550	,097	-,843	-5,649	,000	-,760	-,340

a. Variable dependiente: Log10 (supervivencia)

Aceptando el modelo completo

Para una larva (L1) que pesa 1.58 ¿qué dosis estimamos necesaria para que viva el mismo tiempo que una larva (L2) que pesa 2.51 y a la que se administra una dosis de 3.16?

Solución

Estimación de Log_{10} (Supervivencia) de L2 =

$2.589 + 0.875\ Log_{10}(2.51) - 0.378 Log_{10}(3.16) = 2.75$

Supervivencia estimada de L2 = $10^{2.75} = 562.34$

Dosis estimada para L1

$2.75 = 2.589 + 0.875\ Log_{10}(1.58) - 0.378 Log_{10}(x)$

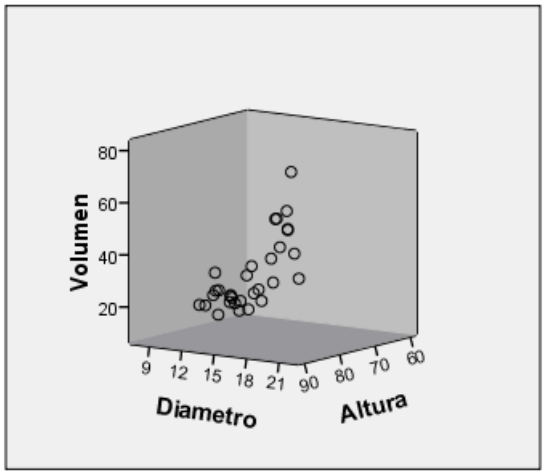
Despejando

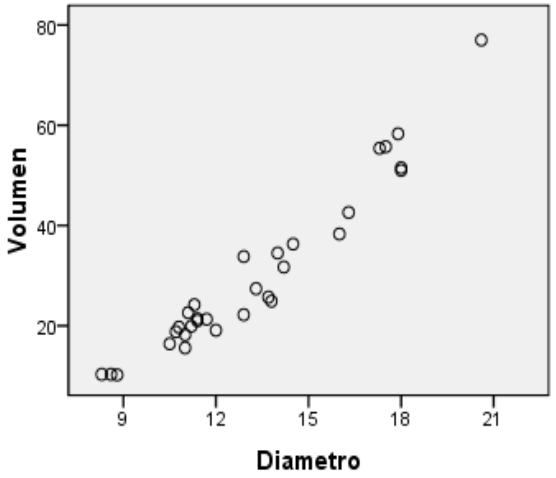
$Log_{10}(x) = 0.04$ → **la dosis pedida es $10^{0.04} = 1.10$**

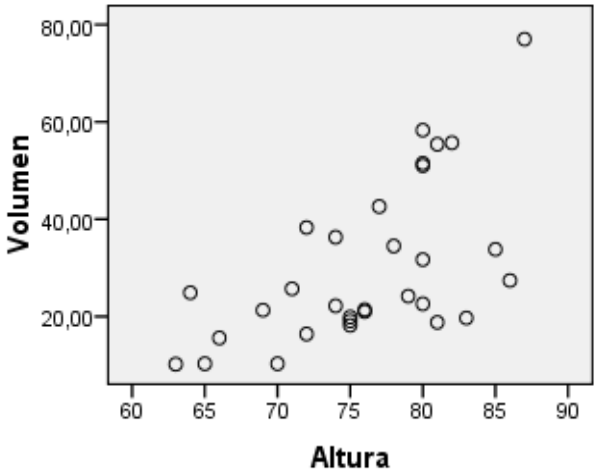
Ejemplo 4

Los siguientes resultados corresponden al análisis realizado sobre los cerezos negros en el Allegheny National Forest, Pennsylvania. Los datos corresponden al volumen (en pies cúbicos), la altura (en pies) y el diámetro (en pulgadas, a 54 pulgadas sobre la base) de 31 cerezos.

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







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			

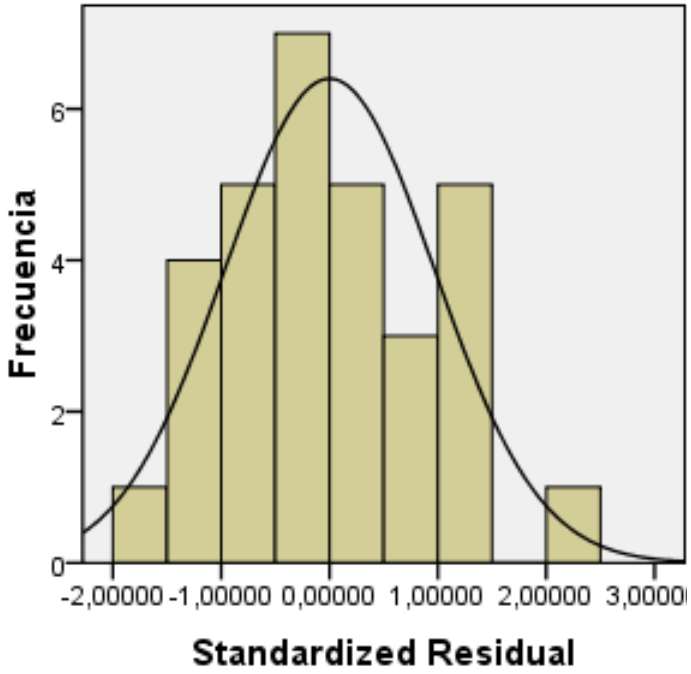
a. Variables predictoras: (Constante), Altura, Diametro

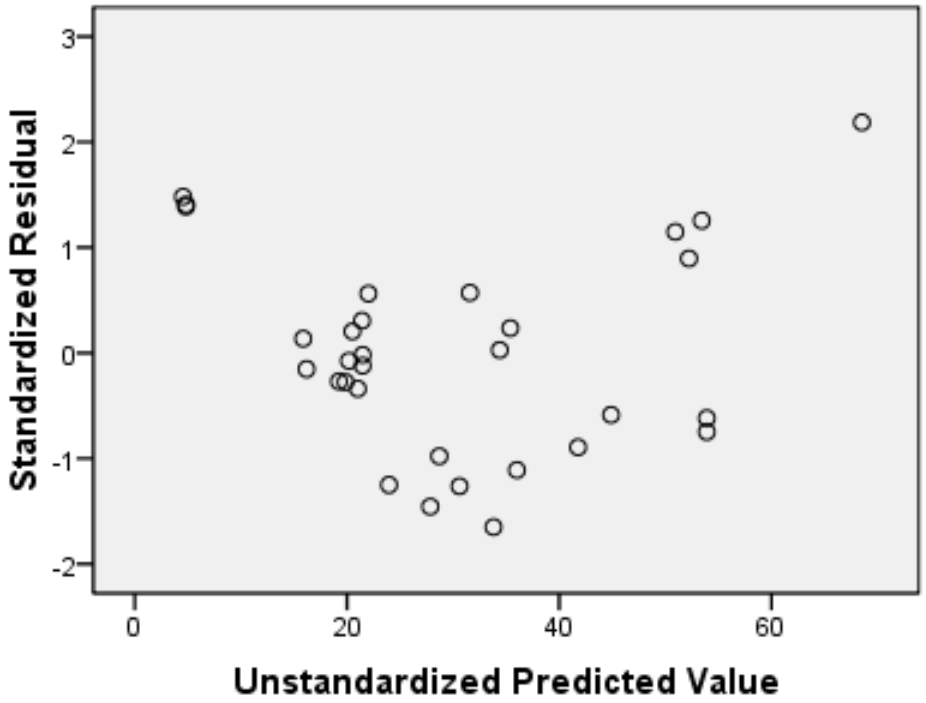
Coeficientes[a]

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

Análisis de los residuos





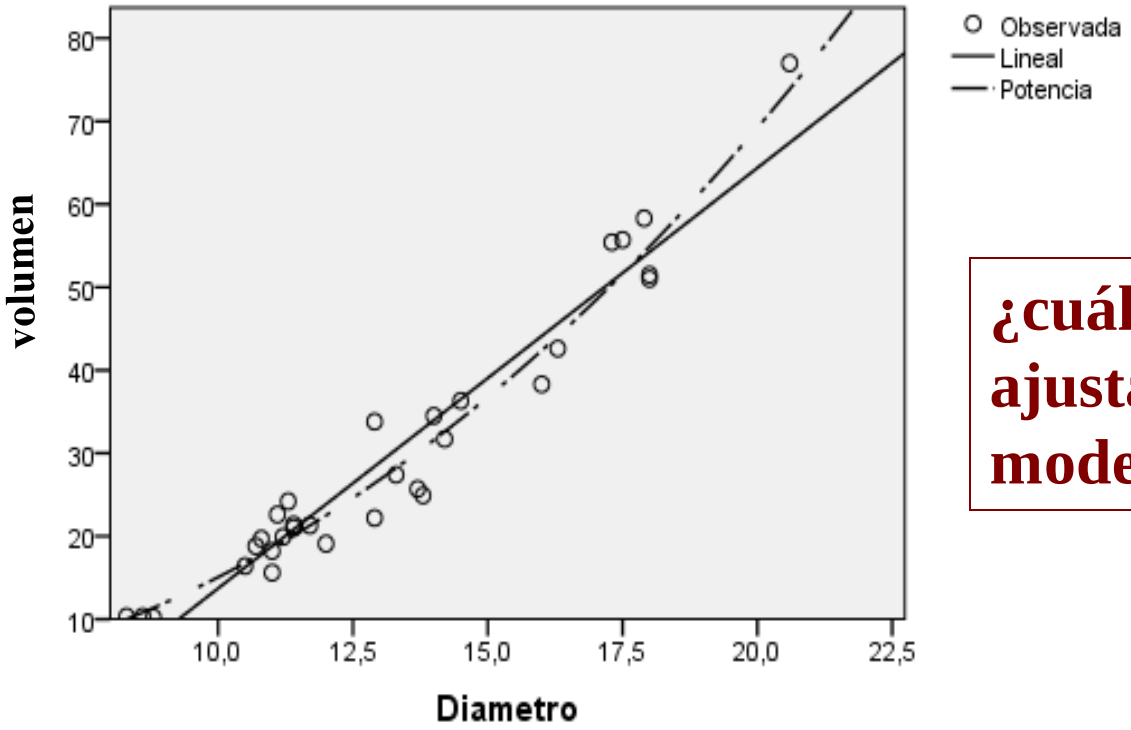
Regresión simple (sólo el diámetro)

Resumen del modelo y estimaciones de los parámetros

Variable dependiente:Volumen

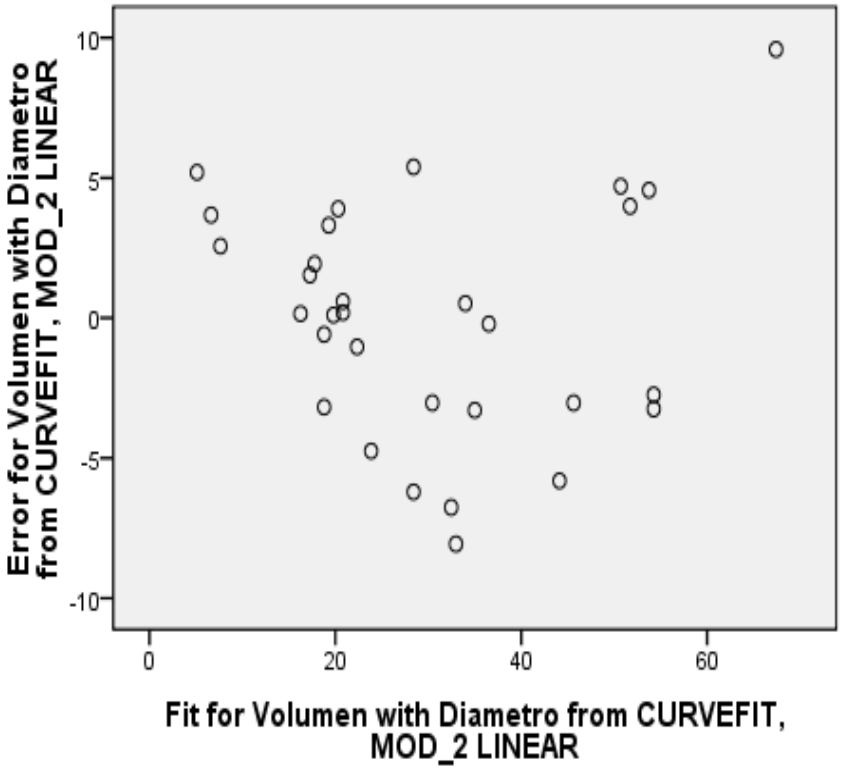
Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,935	419,360	1	29	,000	-36,943	5,066
Potencia	,954	599,717	1	29	,000	,095	2,200

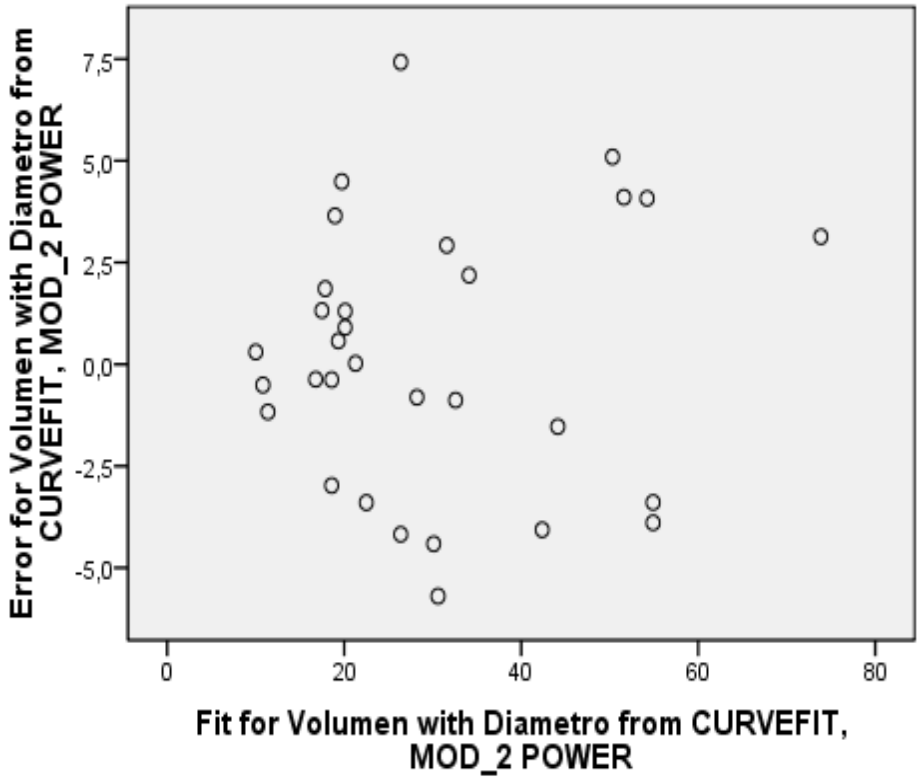
La variable independiente esDiametro.



¿cuál es la curva ajustada con el modelo potencial?

Regresión simple (sólo el diámetro)
Residuos no tipificados





¿justifican los residuos la elección del modelo potencial?

Predicciones puntuales

Para un cerezo con una altura de 80 pies y un diámetro de 16 pulgadas

Con el modelo lineal completo (diámetro y altura):
Volumen estimado = -57,988 + 4,708 (16) +0,339 (80) = 44,46 pies cúbicos

Con el modelo lineal (solo el diámetro):
Volumen estimado = -36,943 + 5,066 (16) = 44,11 pies cúbicos

Con el modelo potencial (sólo el diámetro):
Volumen estimado = $0{,}095\ (16)^{2{,}2}$ = 42,34 pies cúbicos