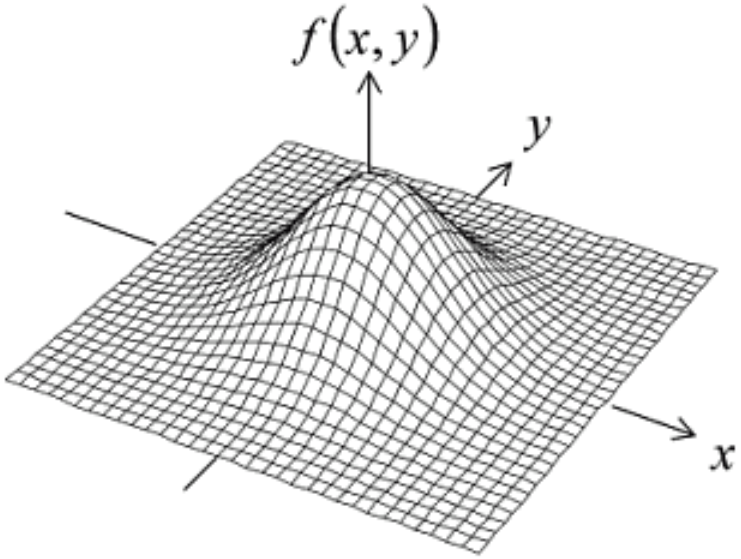
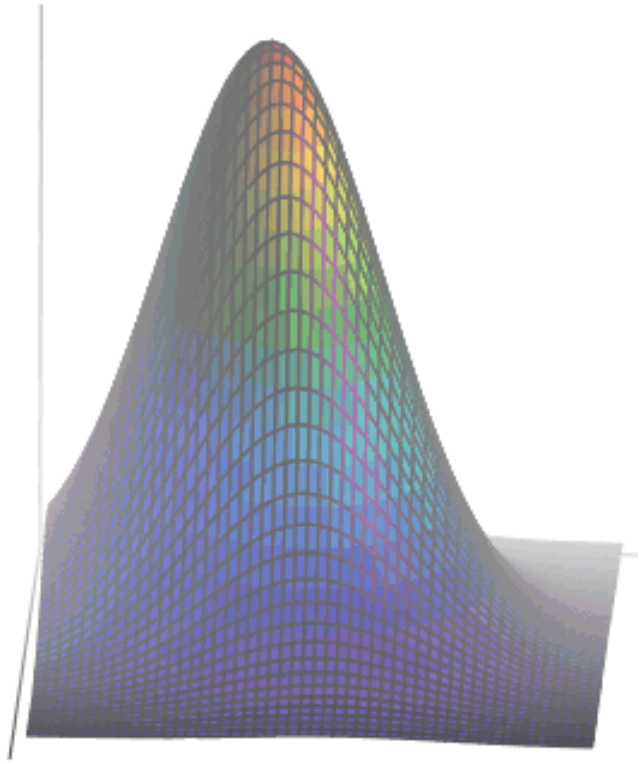
REGRESIÓN LINEAL SIMPLE
Nuevos elementos

- La Normal bivariante
  (modelo de probabilidad)
- Ajuste de una recta a una nube de puntos (análisis de datos)

Distribución Normal Bivariante
(parámetros μ_1, μ_2, σ_1, σ_2, ρ)

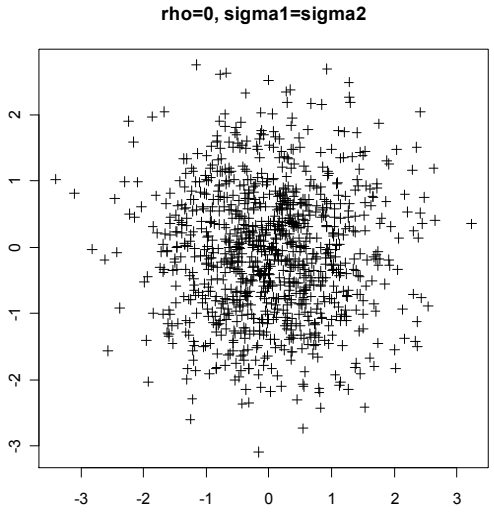




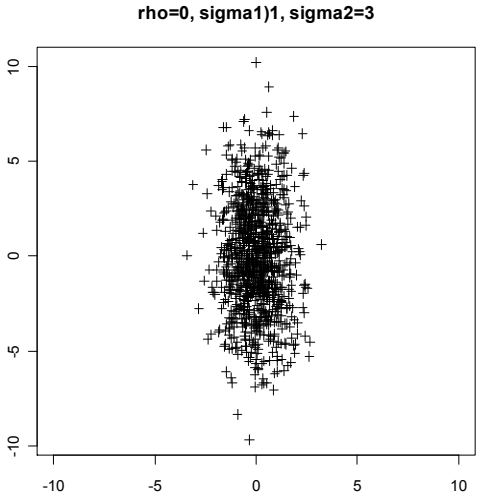
$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2^2(x-\mu_1)^2+\sigma_1^2(y-\mu_2)^2-2\sigma_1\sigma_2\rho(x-\mu_1)(y-\mu_2))\}$$

$\mu_1 = E(X)$ $\mu_2 = E(Y)$ $\sigma_1^2 = Var(X)$ $\sigma_2^2 = Var(Y)$ $\rho =$ **Coef. Correlación** (X,Y)

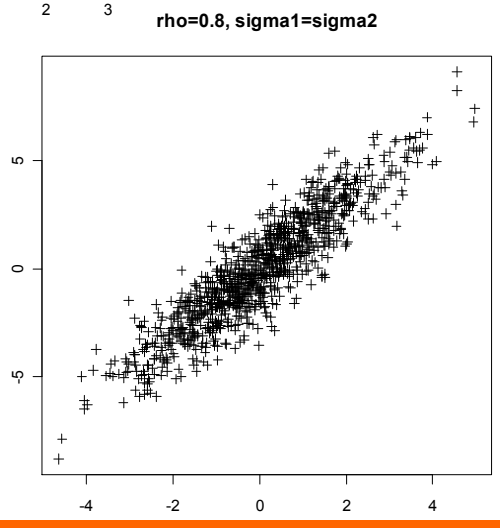
Distribución Normal Bivariante (simulación de datos)



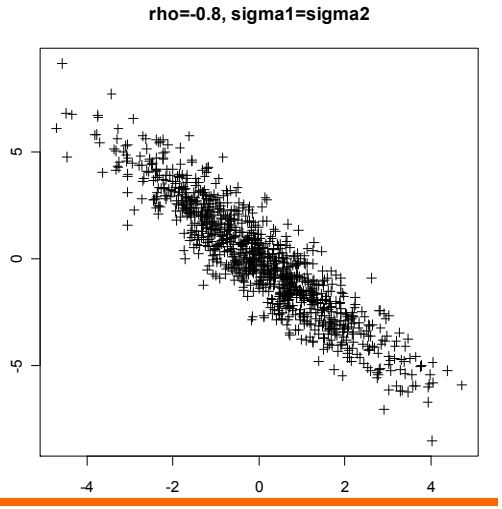
$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = 0$



$\mu_1 = \mu_2 = 0$
$\sigma_1 = 1\ \sigma_2 = 3$
$\rho = 0$

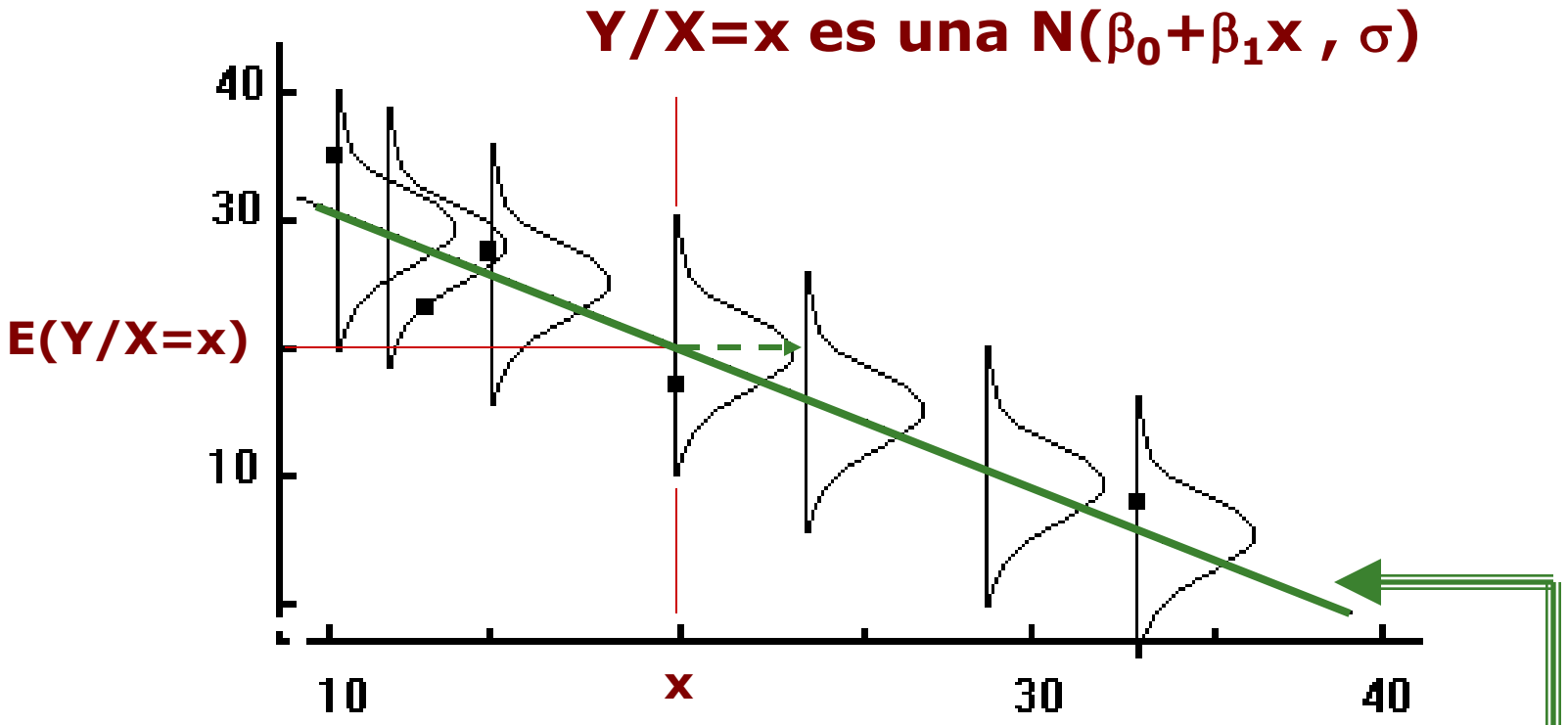


$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = 0.8$



$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = -0.8$

Normal bivariante: Distribuciones condicionadas



Ejemplo de David W. Stockburger
(Modelo para X resultado de un test, Y errores de producción)

$y = \beta_0 + \beta_1 x$ es la recta de regresión de Y sobre X

Las técnicas de Regresión lineal simple parten de dos variables cuantitativas:

La variable explicativa (x)
La variable respuesta (y)

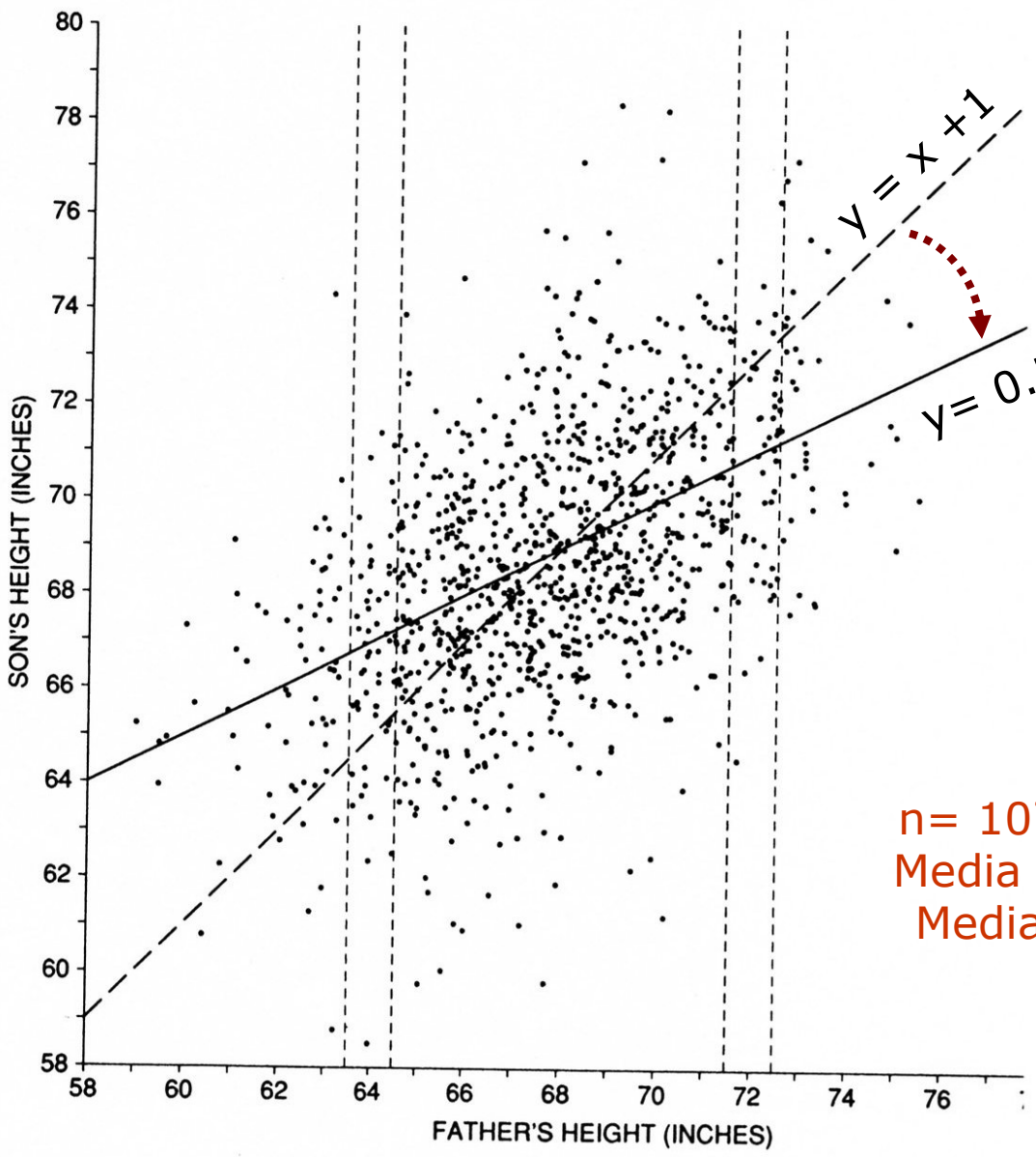
Y tratan de explicar la y mediante una función lineal de la x representada por la recta

$$y = \beta_0 + \beta_1 x$$

Para ello dispondremos:

De un modelo de probabilidad (la Normal) y de n pares de datos (x_i, y_i) que suponemos que provienen del modelo establecido

El origen:On the laws of inheritance in man
Karl Pearson
Biometrika 1903



Variables:
X altura del padre
Y altura del hijo
Datos:
n= 1078 parejas de padres e hijos
Media de los padres = 68 pulgadas
Media de los hijos = 69 pulgadas
$v_x = v_y = 2.7$
r =0.51

Modelo 1 (una variable aleatoria, fijado x)

$$Y = \beta_0 + \beta_1 x + U \qquad U \longrightarrow N(0, \sigma)$$

Modelo 2 (dos variables aleatorias)

$$(X, Y) \longrightarrow N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$$

$$Y/X = x \longrightarrow N(\beta_0 + \beta_1 x, \sigma_2\sqrt{1-\rho^2})$$

σ

¡La diferencia está en cómo se tomarán los datos !

Modelo 1:

El experimentador fija los valores de las x_i y obtiene "al azar" los correspondientes y_i

Modelo 2:

El experimentador obtiene "al azar" parejas de valores (x_i, y_i)

En ambos casos

Los datos son un conjunto de n parejas (x_i, y_i)

Muestra aleatoria

$$Y_i = \beta_0 + \beta_1 x_i + U_i \qquad U_i \longrightarrow N(0, \sigma)$$

Normalidad:	$u_i \sim$ Normal
Linealidad:	$E(u_i) = 0$
Homocedasticidad:	$V(u_i) = \sigma^2$
Independencia:	Los u_i son independientes

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \text{ independientes}$$

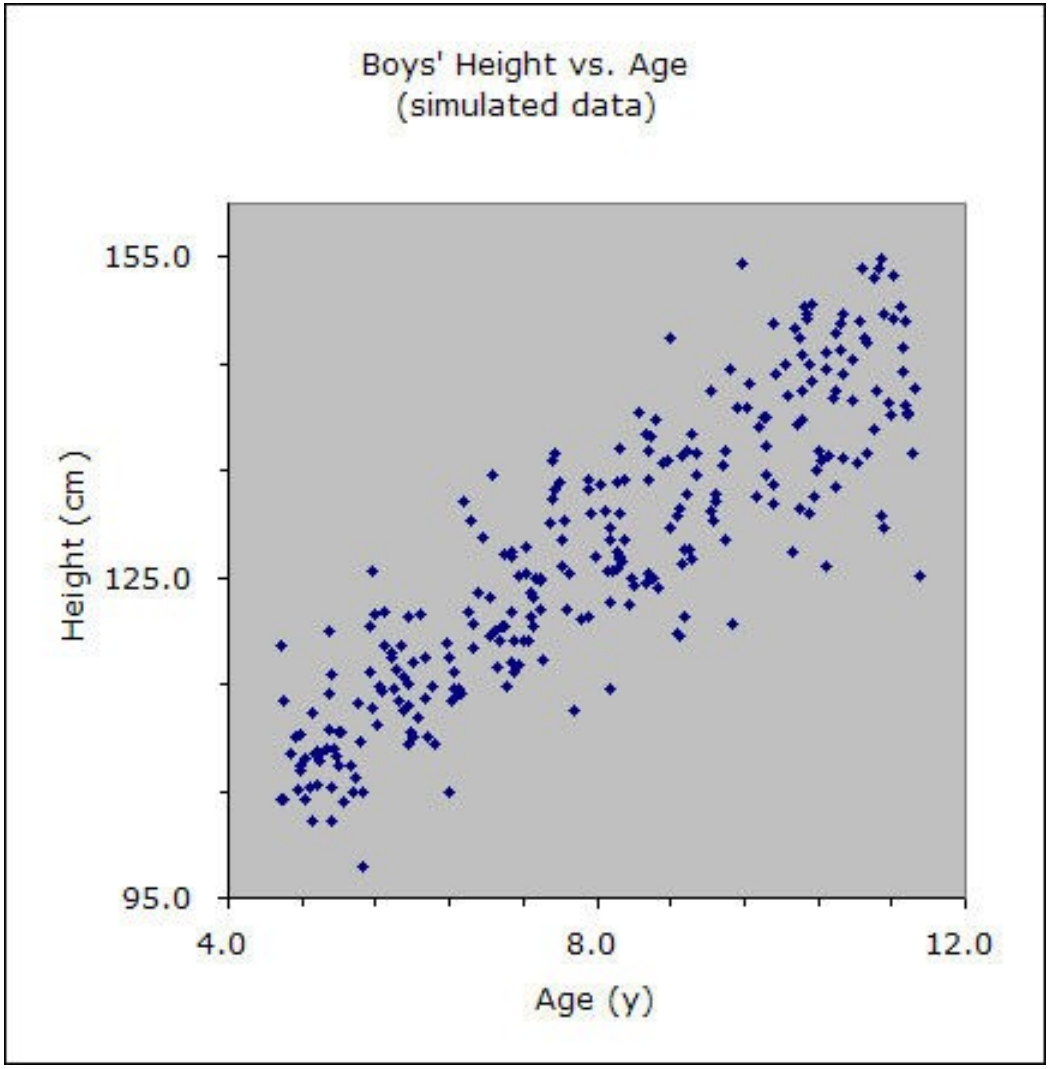
i = 1,2,…,n

Interpretación de los parámetros:

β_0 Representa el valor medio de la respuesta (y) cuando la variable explicativa (x) vale cero

β_1 Representa el incremento de la respuesta media (y) cuando la variable explicativa (x) aumenta en una unidad

Ajuste de una recta a n pares de datos (x_i,y_i)



DATOS

Gráfico de los puntos (x_i,y_i)
$i = 1,2,\ldots,n$

Ajuste de una recta a n pares de datos (x_i, y_i)

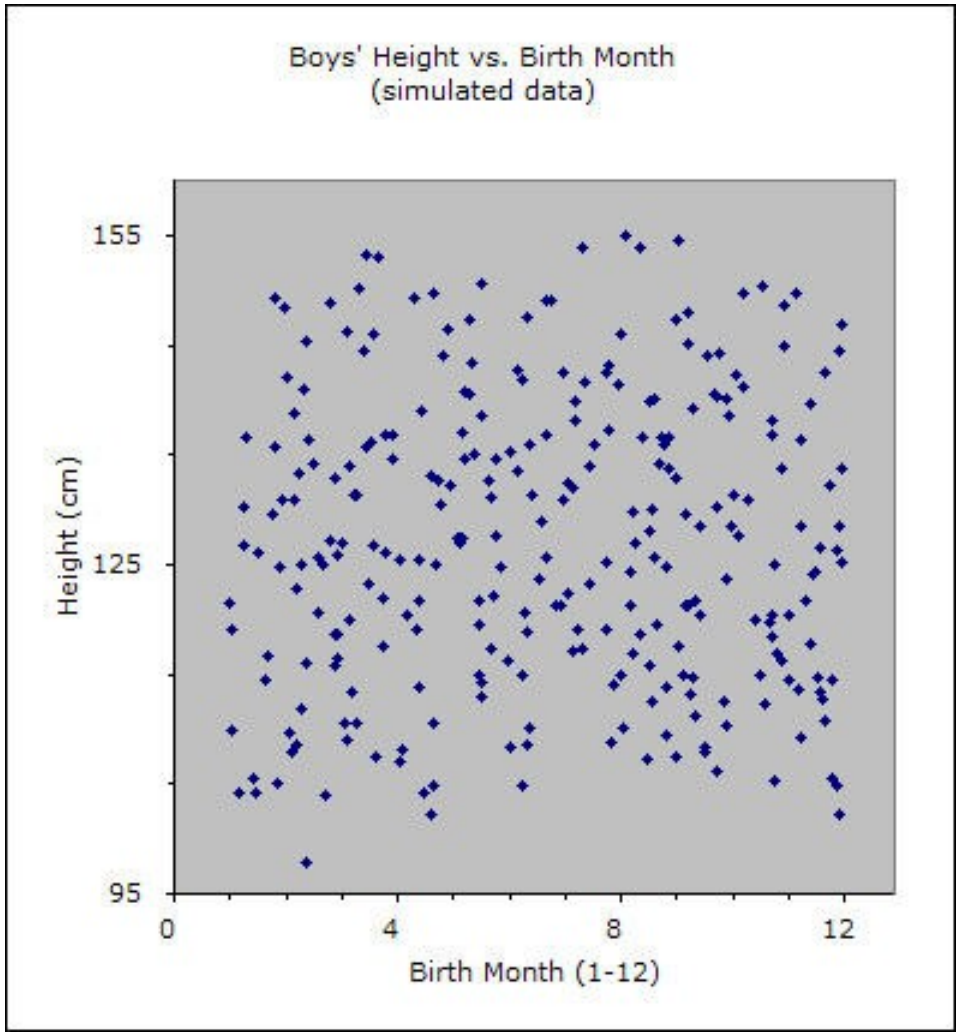
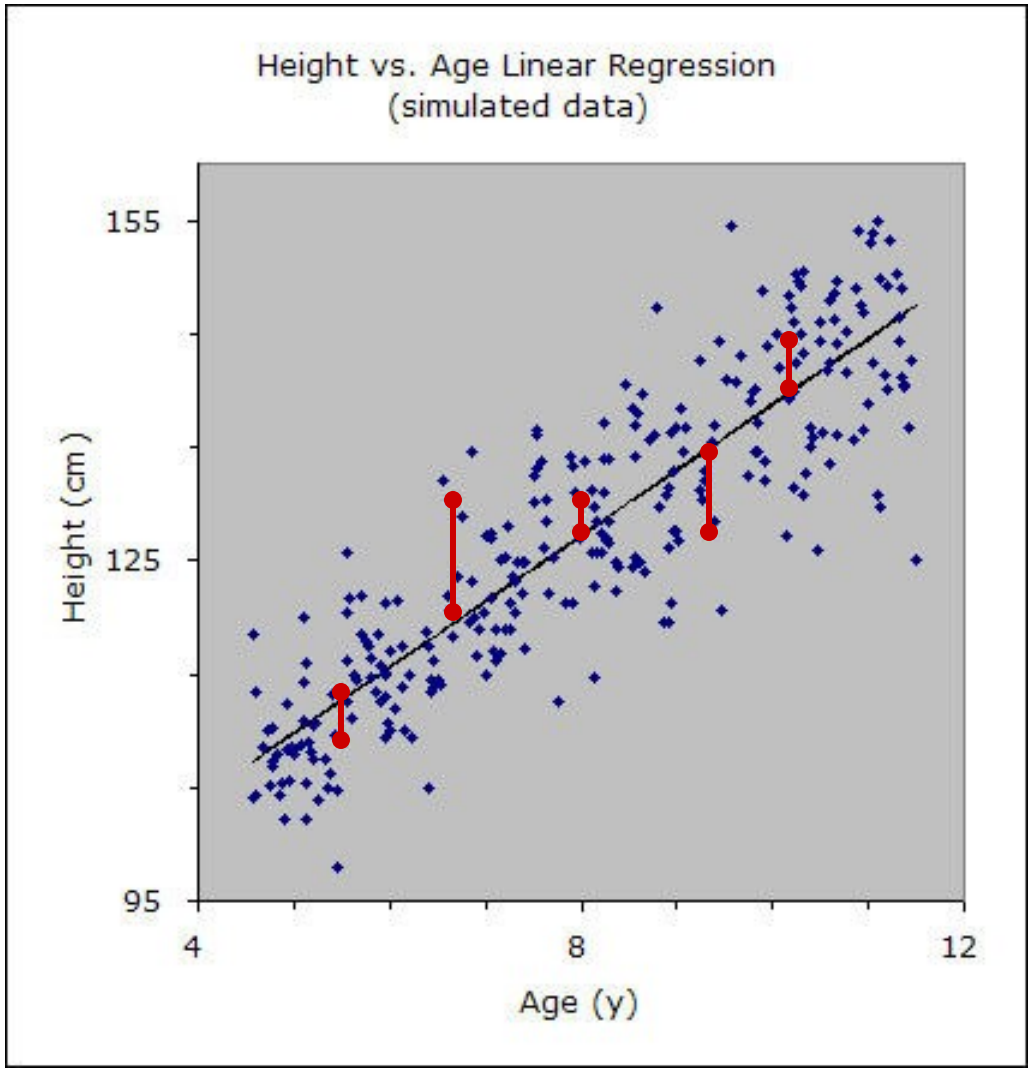


Gráfico de puntos

¿tiene sentido una relación lineal?

¿tiene sentido alguna relación?

Ajuste de una recta a n pares de datos (x_i, y_i)



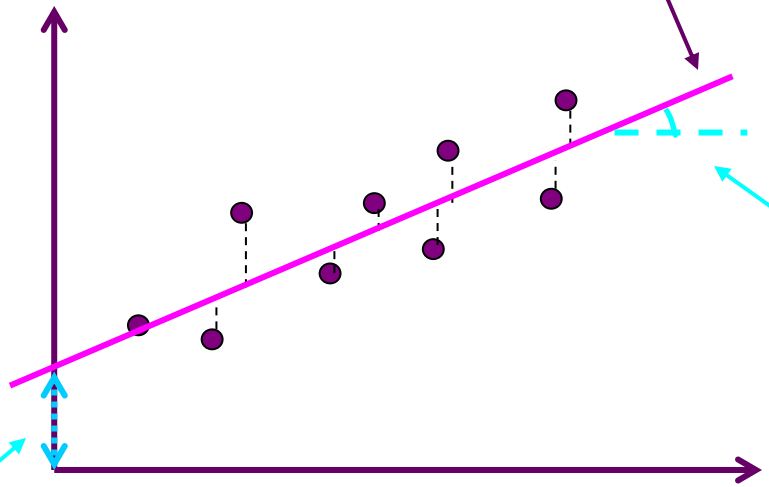
¿Cuál es la recta que mejor predice la altura en función de la edad?

Mínimos cuadrados

Hacemos mínima la suma de los cuadrados de las diferencias entre el valor real de cada y_i con el valor que predice la recta

Ajuste de una recta a n pares de datos (x_i, y_i) Estimación de los coeficientes de la recta

Recta de regresión estimada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

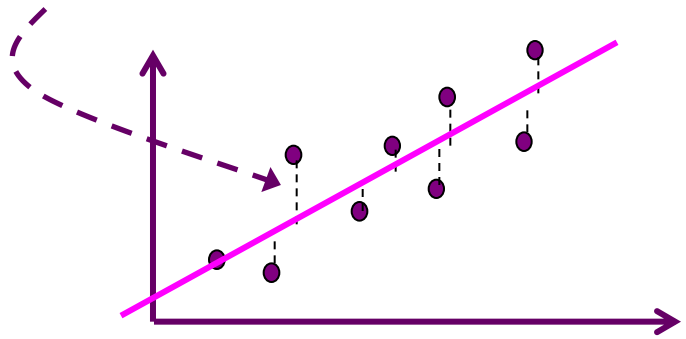


$$\hat{\beta}_1 = \frac{\mathrm{cov}(x_i, y_i)}{\mathrm{var}(x_i)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimación de la varianza residual σ^2
(mide la dispersión de los puntos a la recta)

Los residuos del modelo son



$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$= y_i - \hat{y}_i$$

Los grados de libertad de los residuos son ***n-2***

Varianza residual $S_R^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$

ESTIMACIÓN PUNTUAL DE LOS PARÁMETROS DE LA REGRESIÓN

$$\hat{\beta}_0 = \bar{y} - \frac{cov}{v_x}\bar{x}$$

$$\hat{\beta}_1 = \frac{cov}{v_x}$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Estimación de ρ

$$r = \frac{cov}{\sqrt{v_x v_y}}$$

ESTIMACIÓN POR INTERVALOS DE LOS PARÁMETROS DE LA REGRESIÓN (suponiendo Normalidad)

$$IC_{1-\alpha}(\beta_0) = \left(\hat{\beta}_0 \pm t_{n-2;\alpha/2}\ S_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n v_x}} \right)$$

$$IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{n-2;\alpha/2}\ S_R \sqrt{\frac{1}{n v_x}} \right)$$

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-2)S_R^2}{\chi^2_{n-2;\alpha/2}} ; \frac{(n-2)S_R^2}{\chi^2_{n-2;1-\alpha/2}} \right)$$

Análisis estadístico: requisitos previos

1. **Normalidad:** los datos obtenidos se ajustan razonablemente a una distribución Normal
2. **Homocedasticidad:** la variabilidad de los datos para los distintos valores de x es similar
3. **Linealidad:** los residuos (diferencia de los datos a la recta) se distribuyen sin forma alrededor del cero
4. **Independencia:** las observaciones se realizan de forma independiente unas de otras

SI HAY DESVIACIONES SIGNIFICATIVAS SOBRE ESTOS REQUISITOS LOS RESULTADOS POSTERIORES PUEDEN SER INCORRECTOS

La importancia de los gráficos de puntos
(4 conjuntos de datos emparejados)

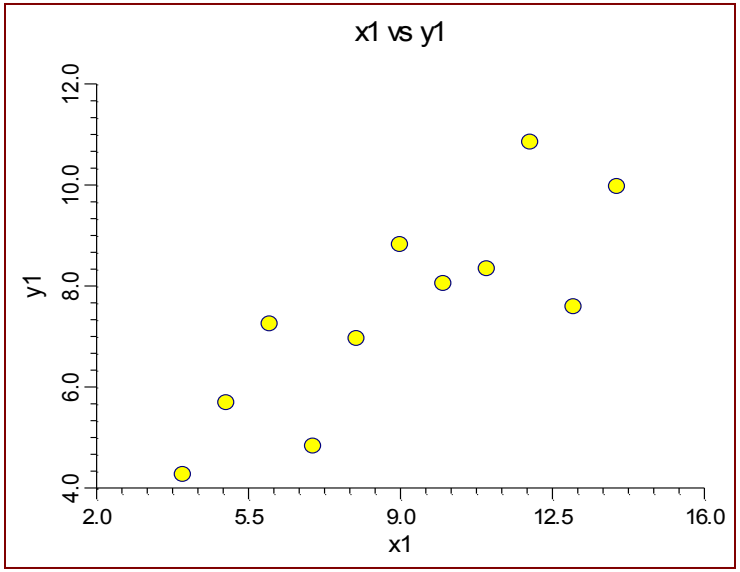
x_1	y_1		x_2	y_2		x_3	y_3		x_4	y_4
10	8.04		10	9.14		10	7.46		8	6.58
8	6.95		8	8.14		8	6.77		8	5.76
13	7.58		13	8.74		13	12.74		8	7.71
9	8.81		9	8.77		9	7.11		8	8.84
11	8.33		11	9.26		11	7.81		8	8.47
14	9.96		14	8.1		14	8.84		8	7.04
6	7.24		6	6.13		6	6.08		8	5.25
4	4.26		4	3.1		4	5.39		19	12.5
12	10.84		12	9.13		12	8.15		8	5.56
7	4.82		7	7.26		7	6.42		8	7.91
5	5.68		5	4.74		5	5.73		8	6.89

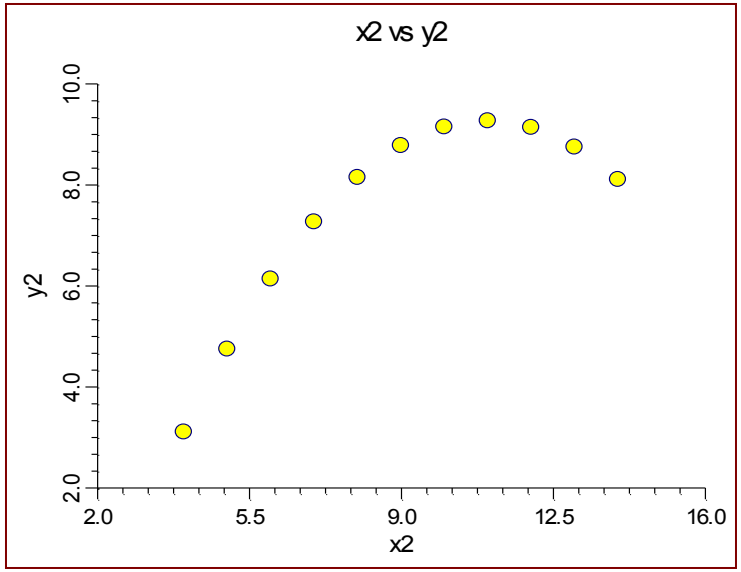
From the Exploring Data website http://curriculum.qed.qld.gov.au/kla/eda/

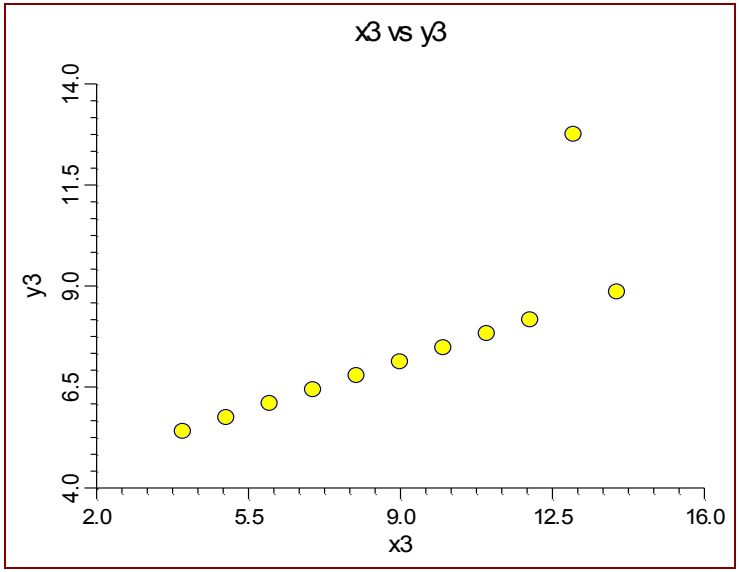
Los 4 grupos de datos tienen exactamente los mismos valores descriptivos siguientes:

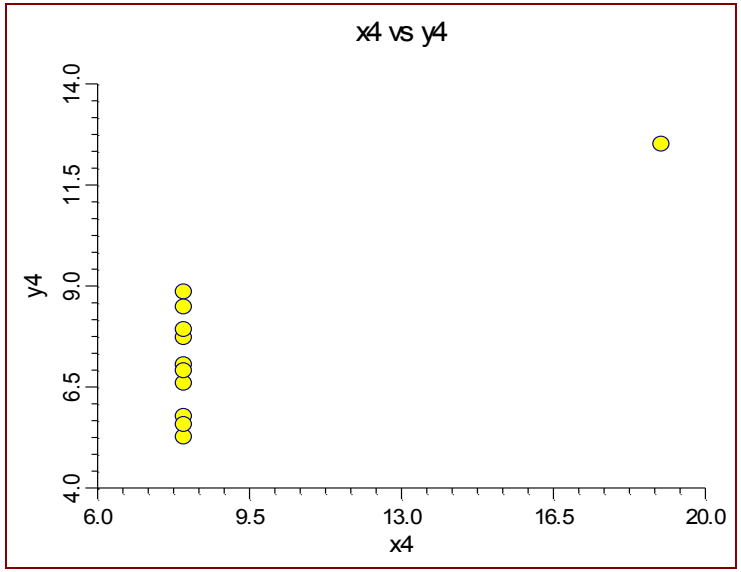
Número de datos	11
Media de las x's	9.0
Media de las y's	7.5
Ecuación de la recta de regresión	y = 3 + 0.5x
Coeficiente de correlación	0.82
r^2	0.67

Pero los gráficos son:









Análisis de los residuos

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

Los residuos pueden dibujarse de distintas formas:

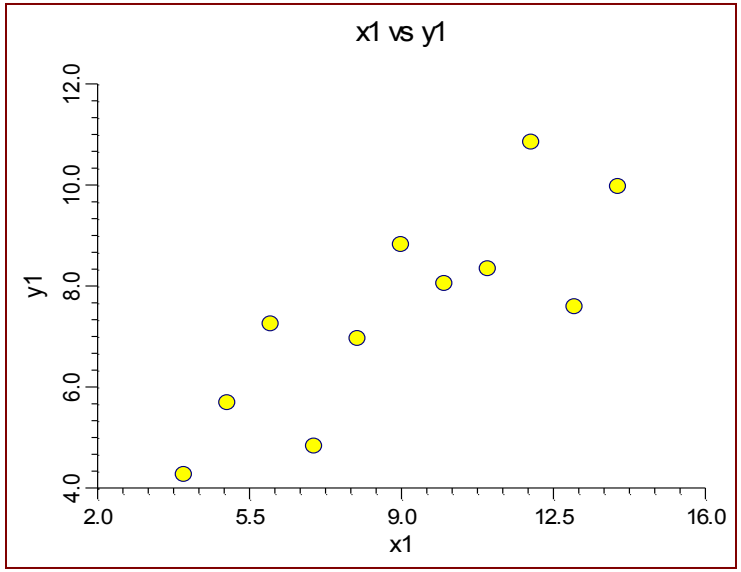
1. poniendo en el eje de abcisas los valores de las x_i y en ordenadas los correspondientes e_i
2. poniendo en el eje de abcisas los valores de las y_i y en ordenadas los correspondientes e_i

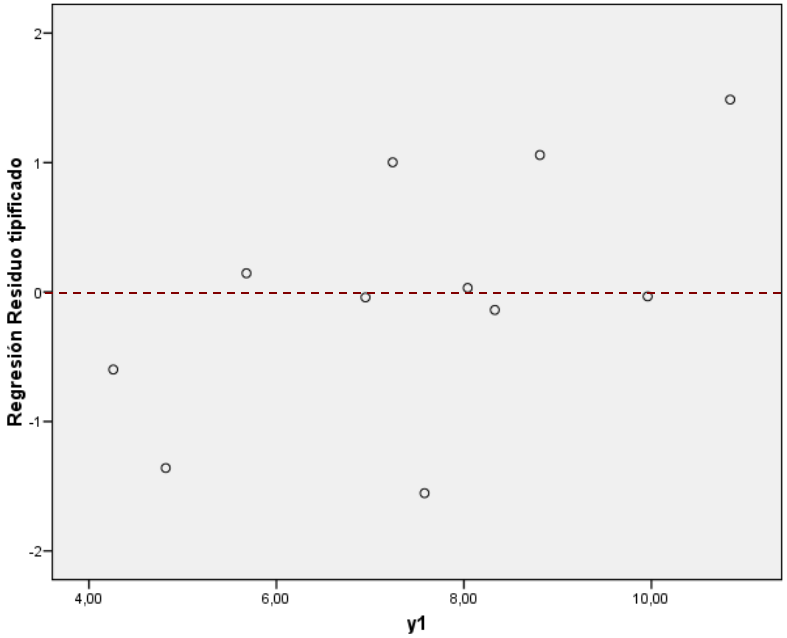
Residuos tipificados o estandarizados

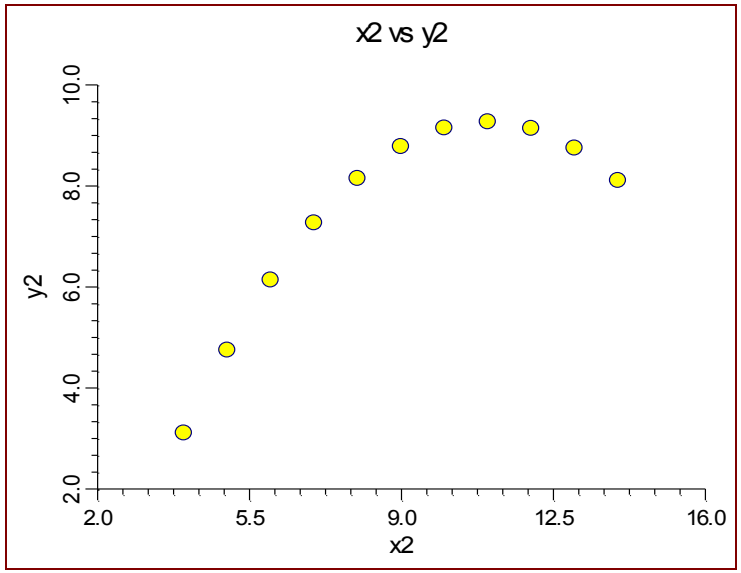
Para evitar la influencia de las unidades de medida utilizadas en los datos y eliminar posibles diferencias debidas al azar en su variabilidad, se utilizan los residuos tipificados dividiendo cada uno de ellos por una medida común de la dispersión.

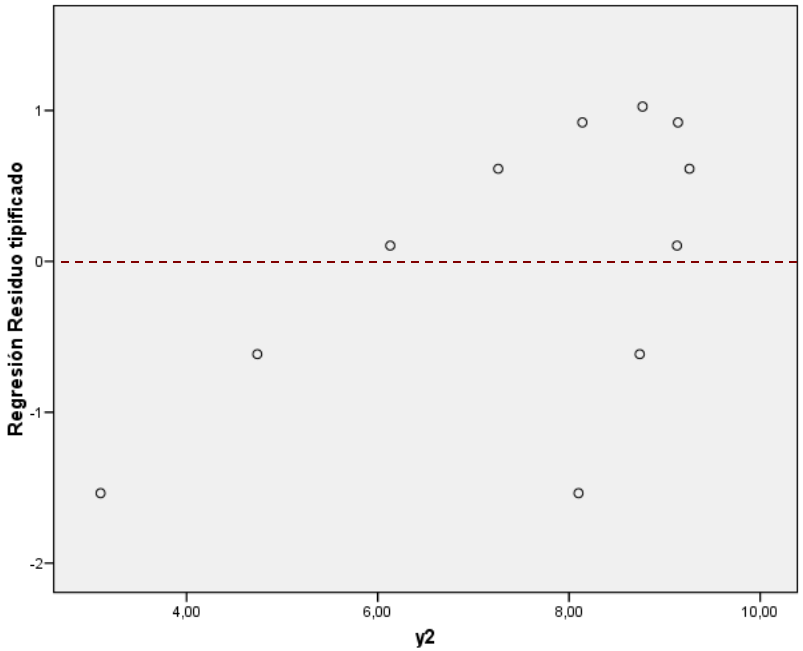
Si el modelo es correcto los residuos tipificados se ajustarán aproximadamente a una N(0,1) y su dispersión será homogénea alrededor del cero.

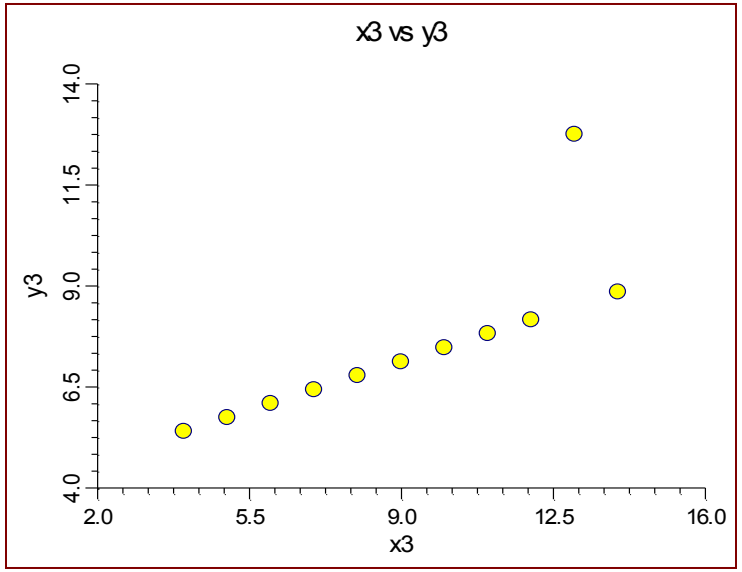
Residuos tipificados muy alejados del cero (fuera de (-2,2)) pueden indicar datos anómalos.

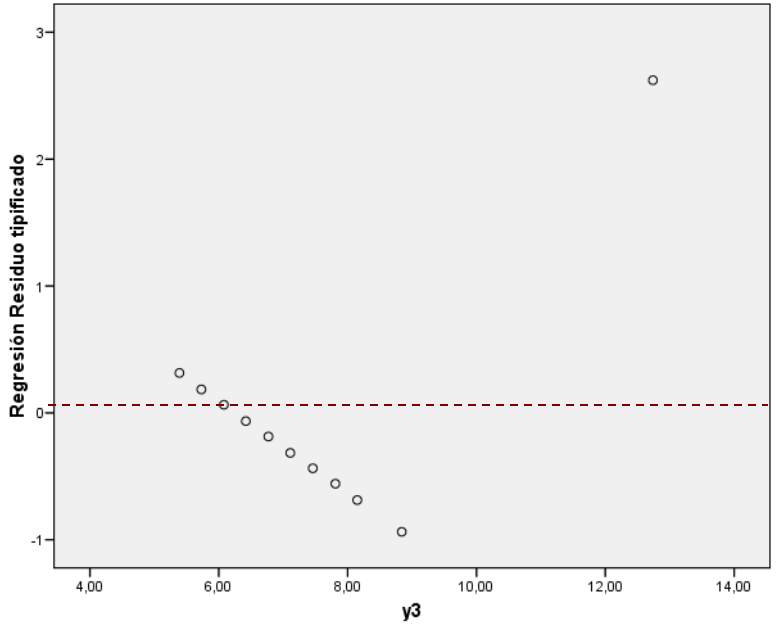

x1 vs y1
y1
12.0
10.0
8.0
6.0
4.0
2.0
5.5
9.0
12.5
16.0
x1

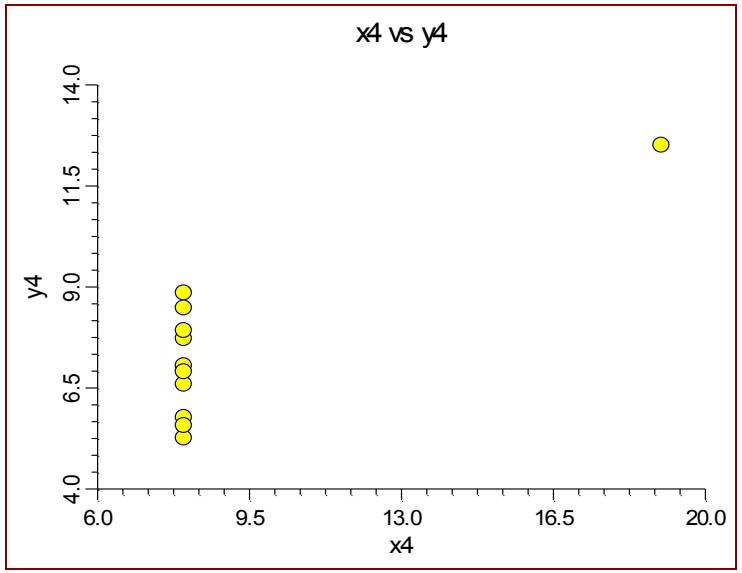

Regresión Residuo tipificado
2
1
0
-1
-2
4,00
6,00
8,00
10,00
y1


x2 vs y2
y2
10.0
8.0
6.0
4.0
2.0
2.0
5.5
9.0
12.5
16.0
x2


Regresión Residuo tipificado
1
0
-1
-2
4,00
6,00
8,00
10,00
y2


x3 vs y3
y3
14.0
11.5
9.0
6.5
4.0
2.0
5.5
9.0
12.5
16.0
x3


Regresión Residuo tipificado
3
2
1
0
-1
4,00
6,00
8,00
10,00
12,00
14,00
y3


x4 vs y4
y4
14.0
11.5
9.0
6.5
4.0
6.0
9.5
13.0
16.5
20.0
x4

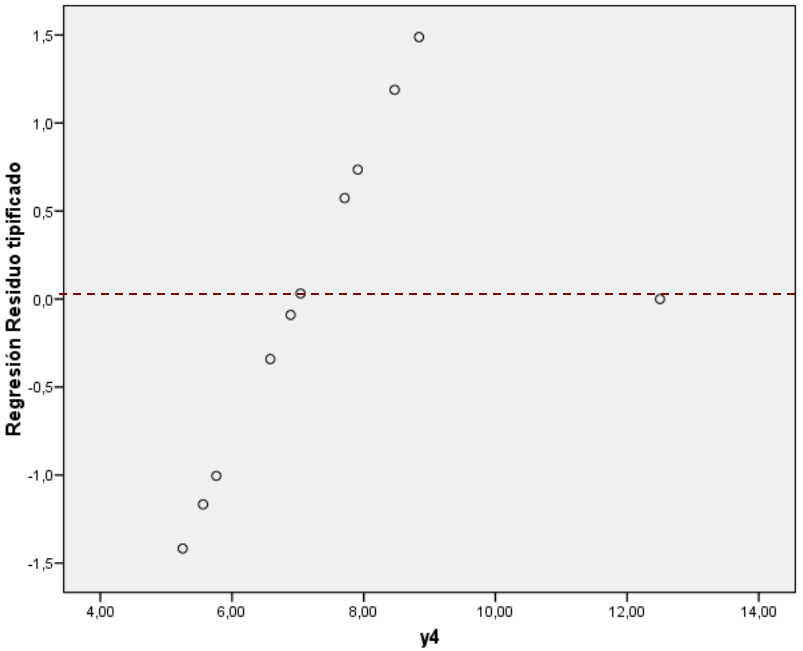
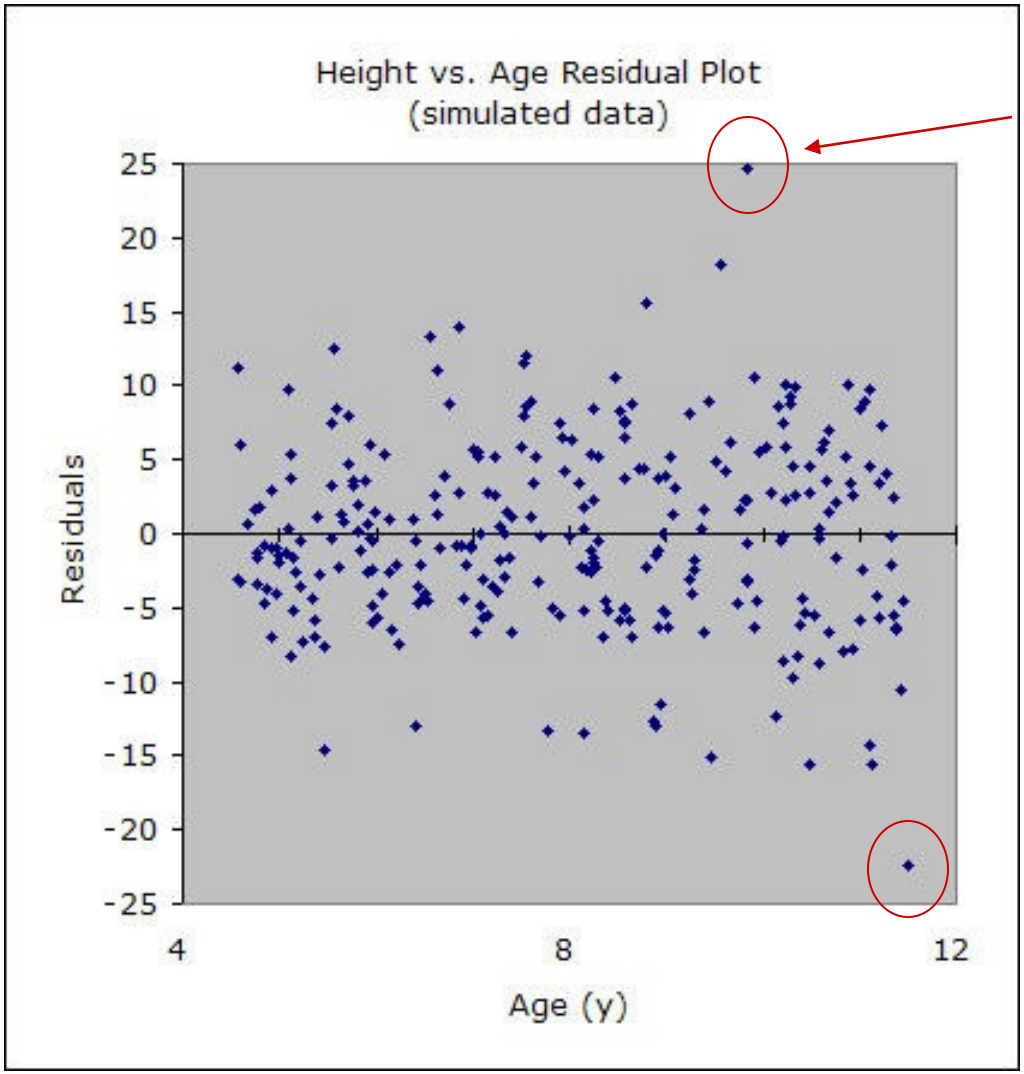

Regresión Residuo tipificado
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
4,00
6,00
8,00
10,00
12,00
14,00
y4

Gráfico de los residuos e_i



¿es este un valor anómalo?

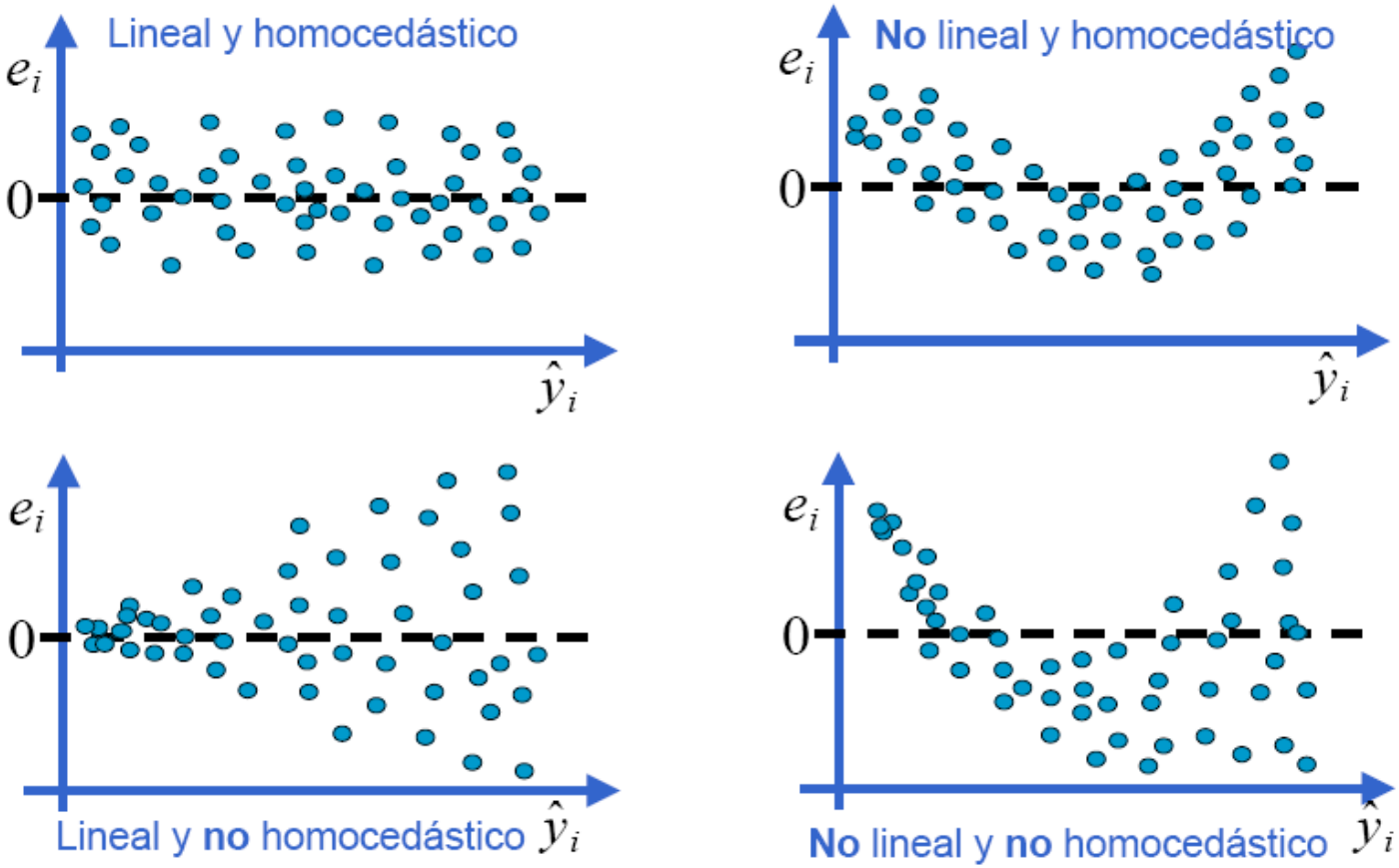
$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

En abcisas los valores de x_i (edades en años)

En ordenadas los residuos e_i sin tipificar

RESIDUOS – VALORES PRONOSTICADOS

¿se cumplen las hipótesis del modelo?

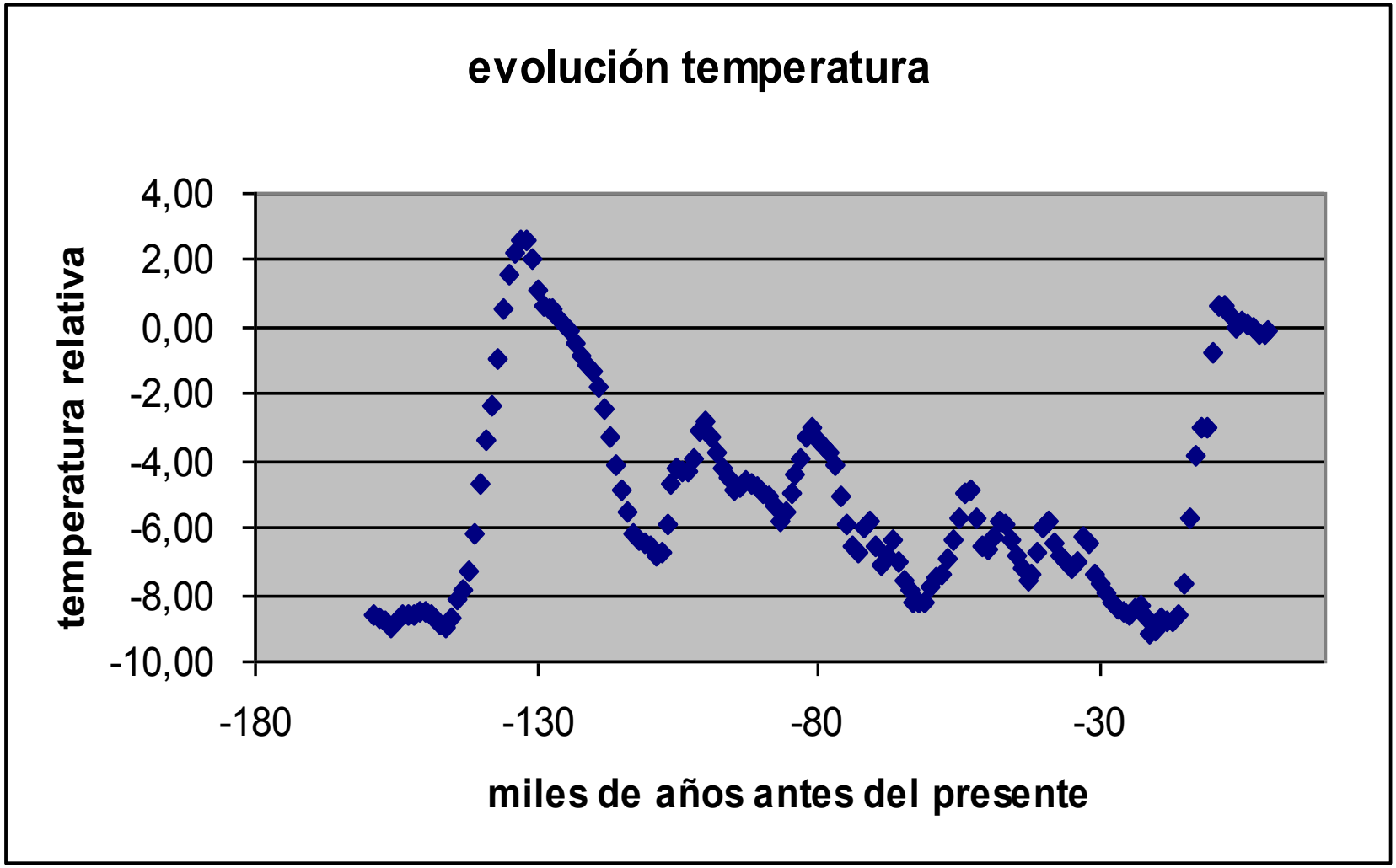


Ejemplo: tiempo, temperatura, CO_2 (desde hace 159.000 años)

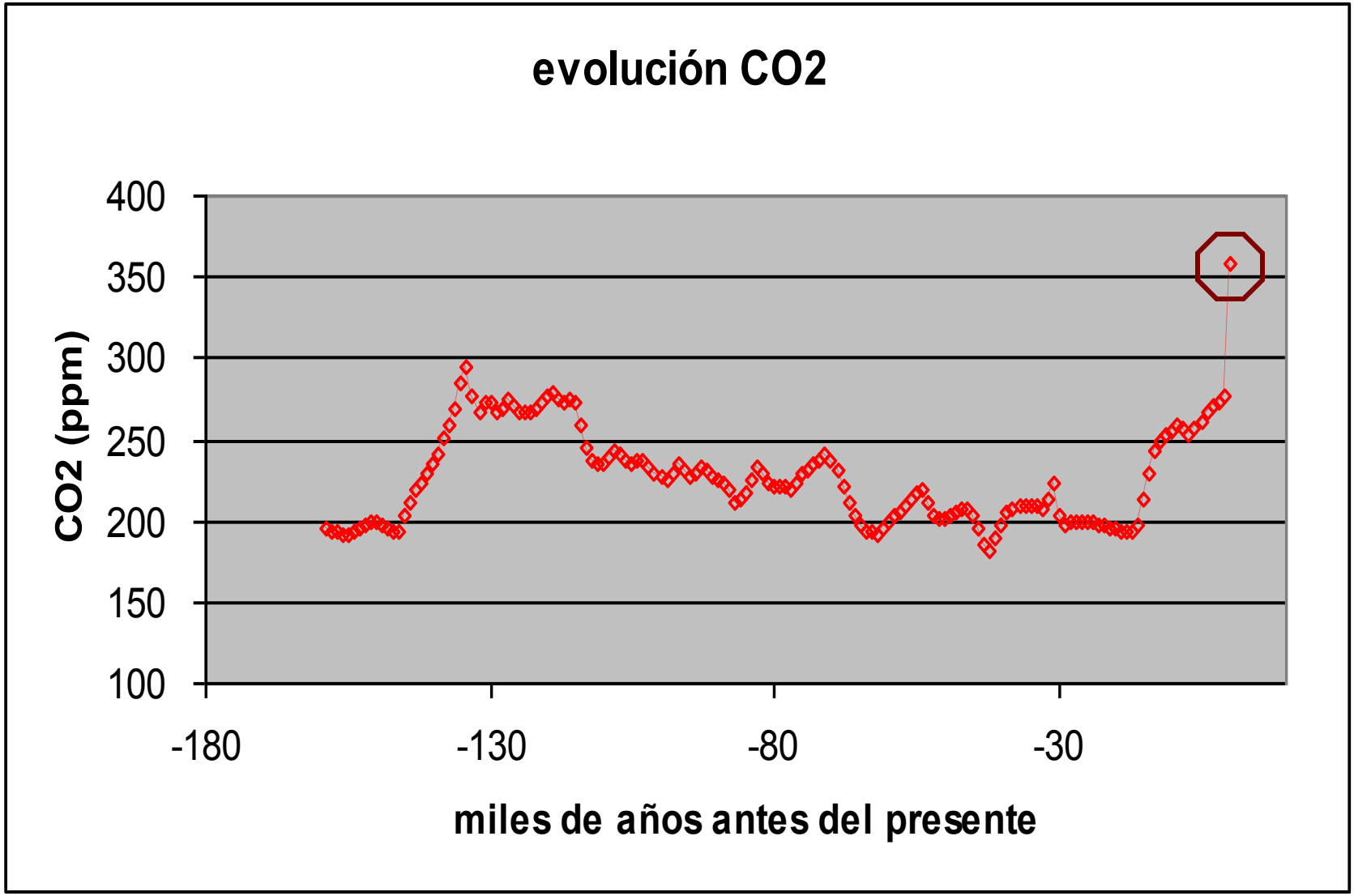
Los siguientes gráficos corresponden a datos obtenidos sobre la evolución de la temperatura global y la concentración atmosférica de CO_2 en los últimos 159.000 años. Las variables son: miles de años antes del presente, diferencia de temperatura respecto a la actual y concentración de CO_2 en la atmósfera.

Source: Compiled by Worldwatch Institute from J.M. Barnola et al. "Historical CO2 Record from the Vostok Ice Core," in Thomas A. Boden et al., eds., Trends '93: A Compendium of Data on Global Change (Oak Ridge, TN.: Oak Ridge National Laboratory, 1994); J. Jouzel et al., "Vostok Isotopic Temperature Record," in Thomas A. Boden et al.; Timothy Whorf, Scripps Institution of Oceanography, La Jolla, CA, private communication, February 2, 1995.

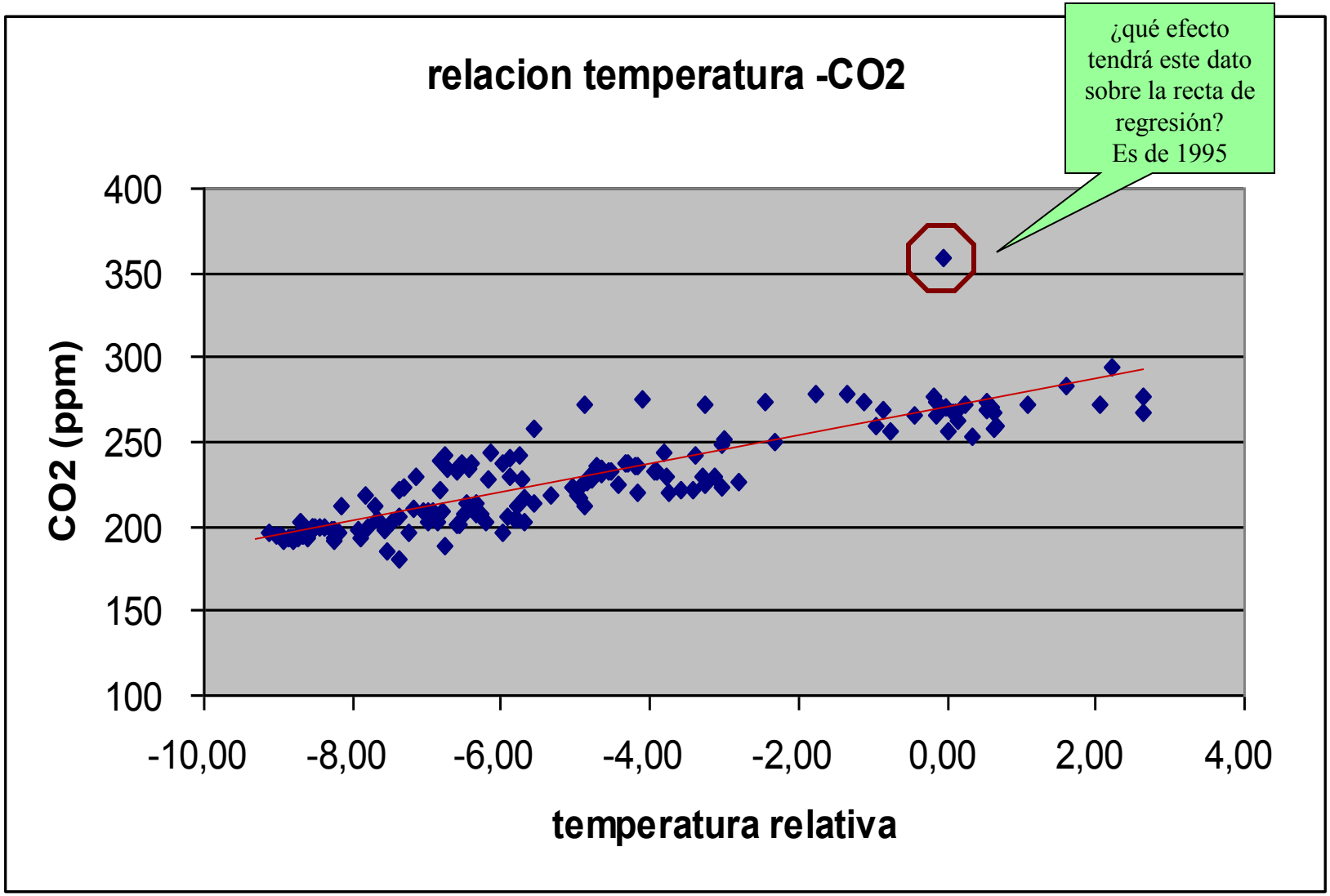
Education Queensland



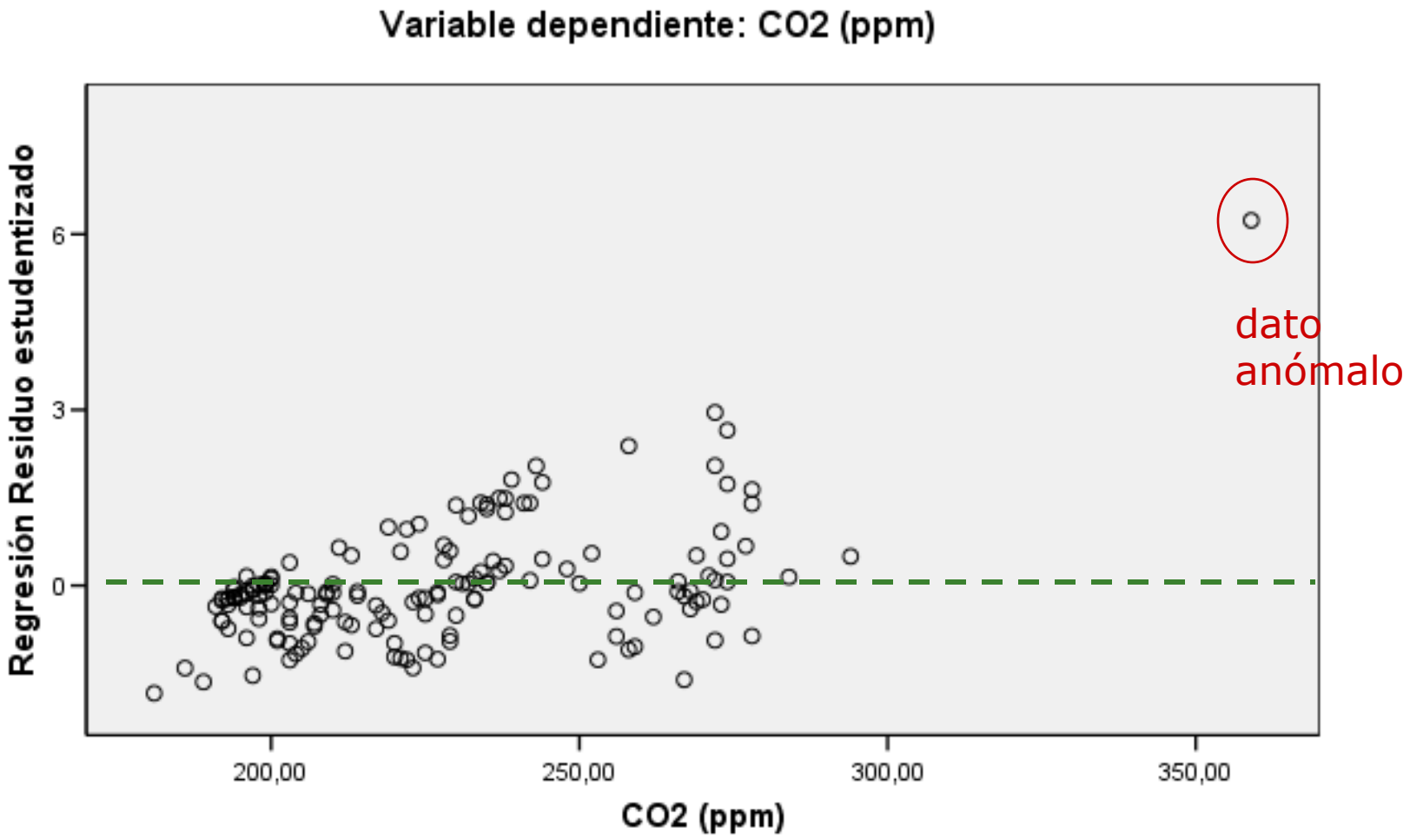
La regresión lineal no es aplicable



La regresión lineal no es aplicable


relacion temperatura -CO2
¿qué efecto tendrá este dato sobre la recta de regresión? Es de 1995
CO2 (ppm)
400
350
300
250
200
150
100
-10,00
-8,00
-6,00
-4,00
-2,00
0,00
2,00
4,00
temperatura relativa

Residuos tipificados sobre la variable dependiente



TRANSFORMACIONES DE LOS DATOS

Cuando detectamos problemas de

no linealidad
o
heterocedasticidad

Y queremos aplicar las técnicas de regresión lineal

Algunas funciones linealizables

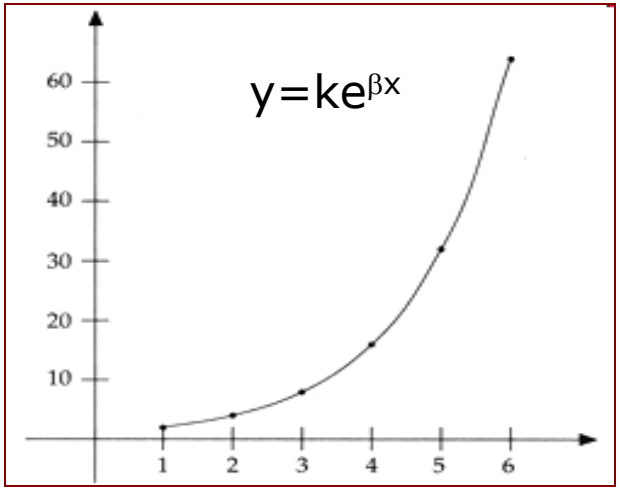
$$y = ke^{\beta x} \xrightarrow{\text{Log}} logy = logk + \beta x$$

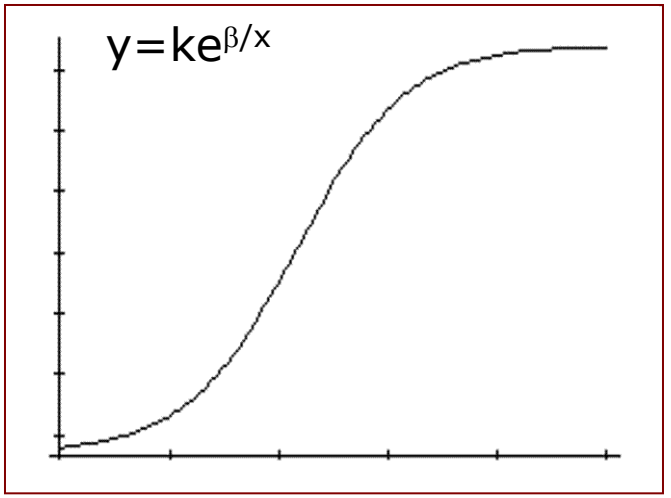
$$y = kx^{\beta} \xrightarrow{\text{Doble Log}} logy = logk + \beta logx$$

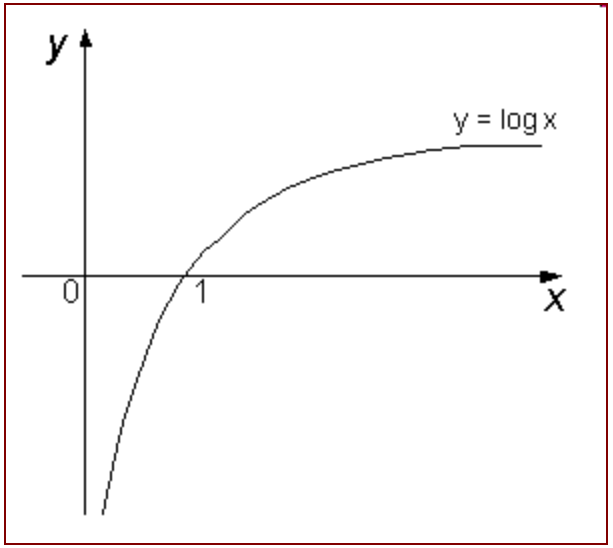
Inversa $\quad y = k + \beta\frac{1}{x}$

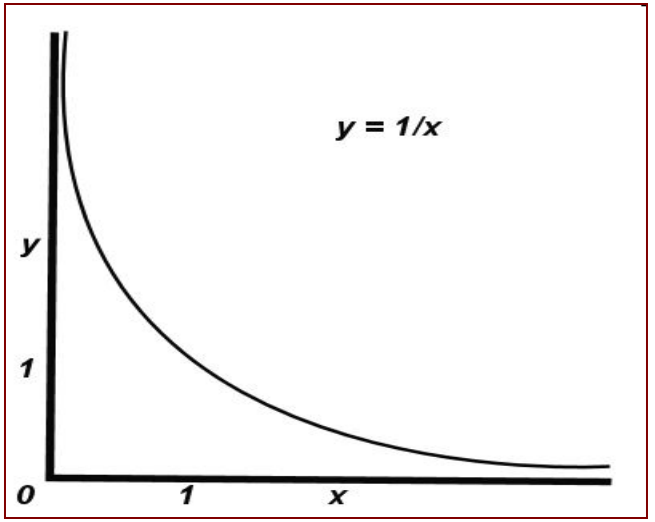
$$y = ke^{\frac{\beta}{x}} \xrightarrow{\text{Log + 1/x}} logy = logk + \beta\frac{1}{x}$$

Algunas gráficas









La curva logística

$$y_i = \frac{C}{1 + e^{-\alpha - \beta X_i}}$$

Nota: C es el valor máximo posible de la variable Y

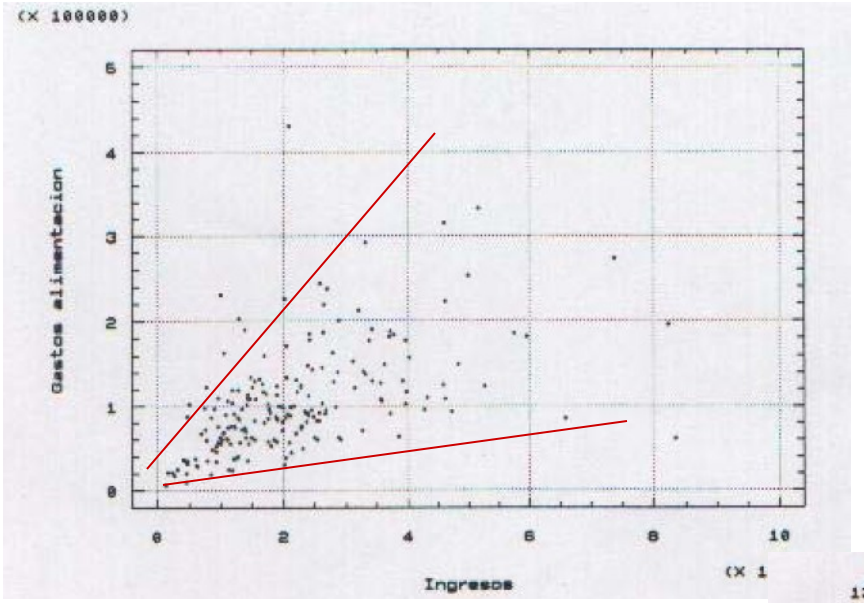
Cambio de variable:

$$Ln\left(\frac{y_i}{(C - y_i)}\right) = Z_i$$

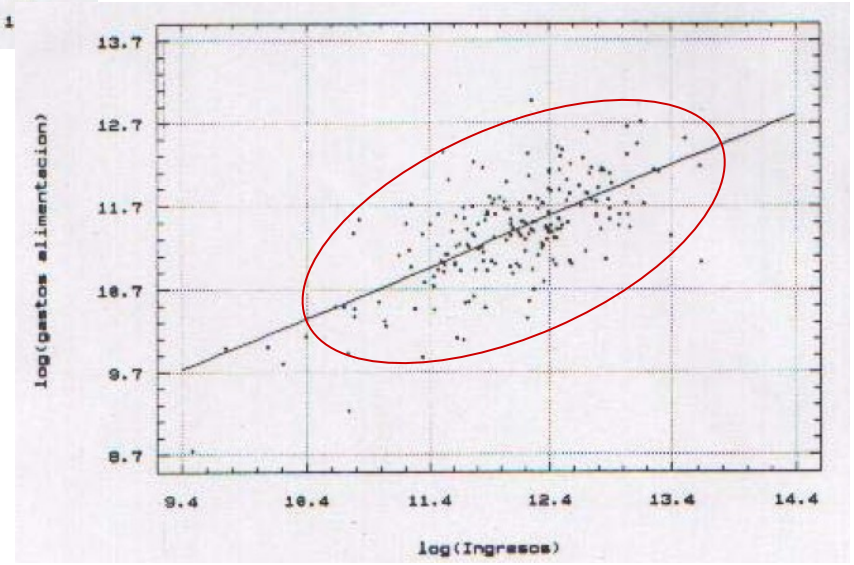
Modelo lineal $\longrightarrow$ $Z_i = \alpha + \beta X_i$

Ejemplo

Problemas de Heterocedasticidad



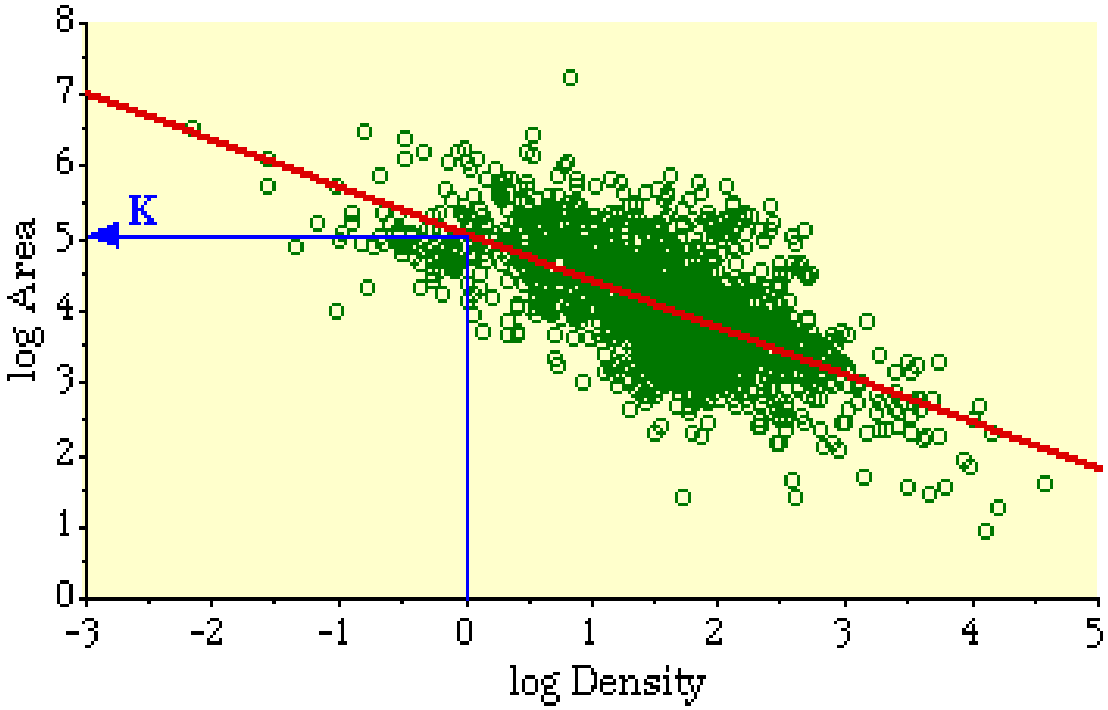
Transformamos



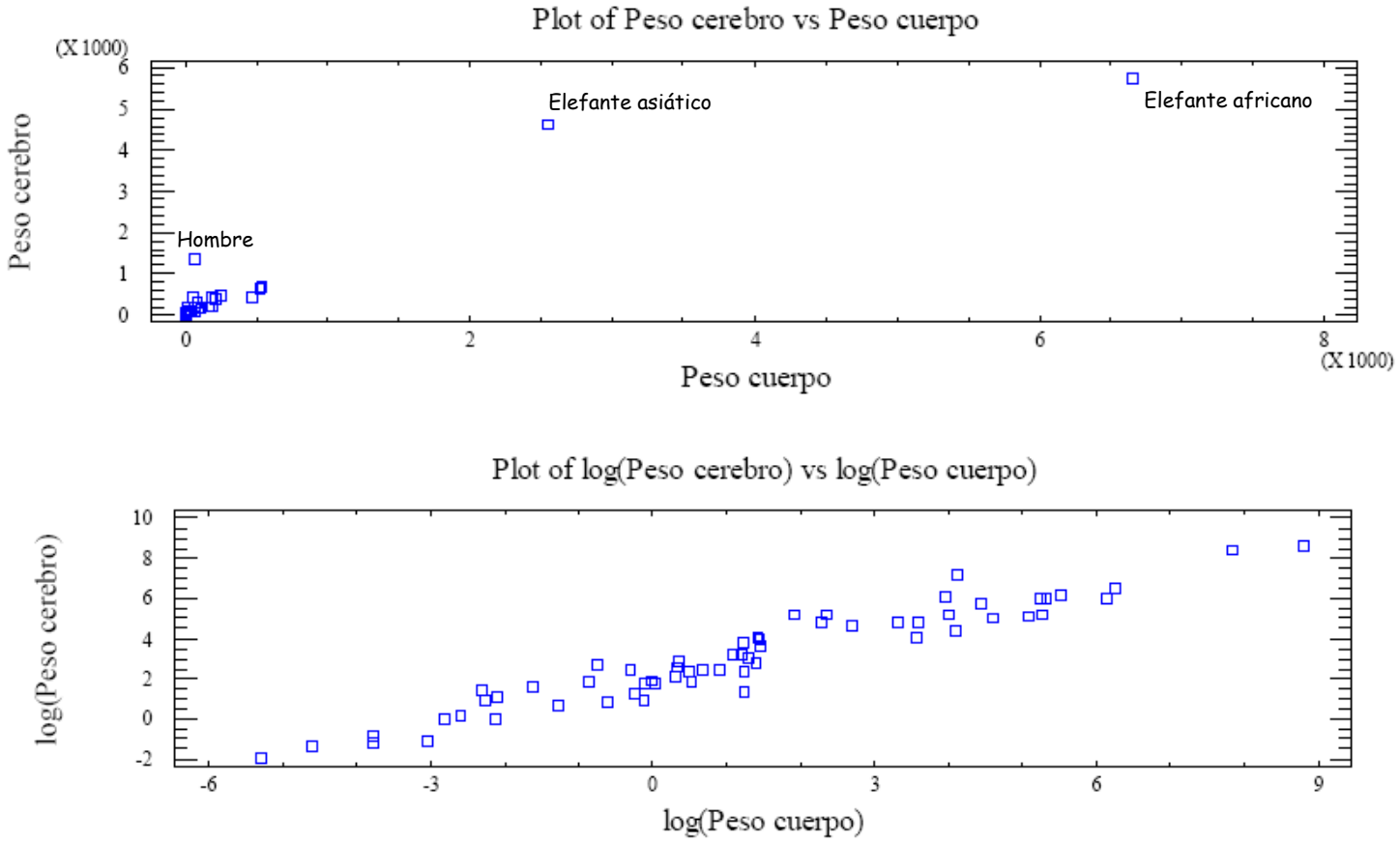
Transformamos

Ejemplo

WORLD REGRESSION LINE
(N = 1,764 primary administrative subdivisions of 98 nations)

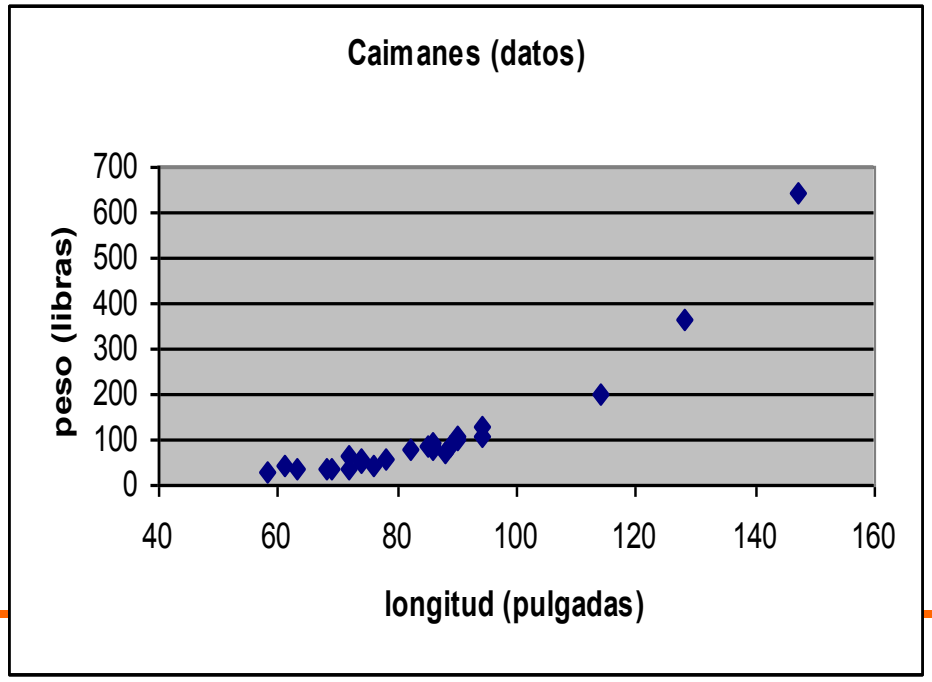


Ejemplo. Peso del cerebro en función del peso corporal para 62 especies de mamíferos



Ejemplo 1. Longitud versus peso

En estudios sobre poblaciones de animales salvajes muchas veces se obtiene información basada en fotografías aéreas. A través de dicha información es posible conocer algunas características de los animales. La longitud de un caimán es fácil de determinar con fotografías aéreas, pero su peso es mucho más difícil de estimar. Para establecer un modelo que estime el peso conocida la longitud del cuerpo, se capturaron 25 caimanes en Florida, midiendo en cada uno su longitud y su peso (Education Queensland, 1997). Los resultados se muestran en la siguiente gráfica:

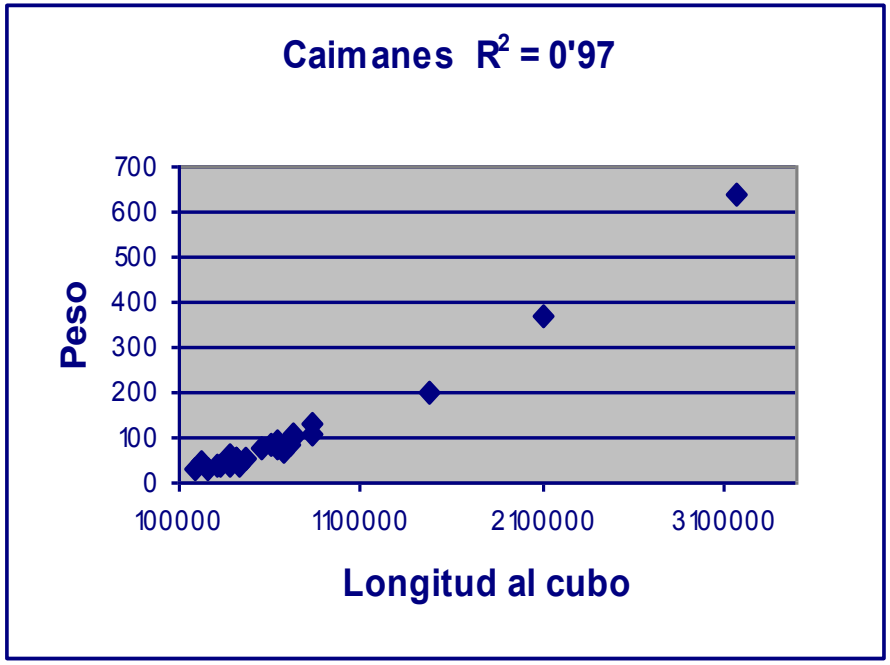


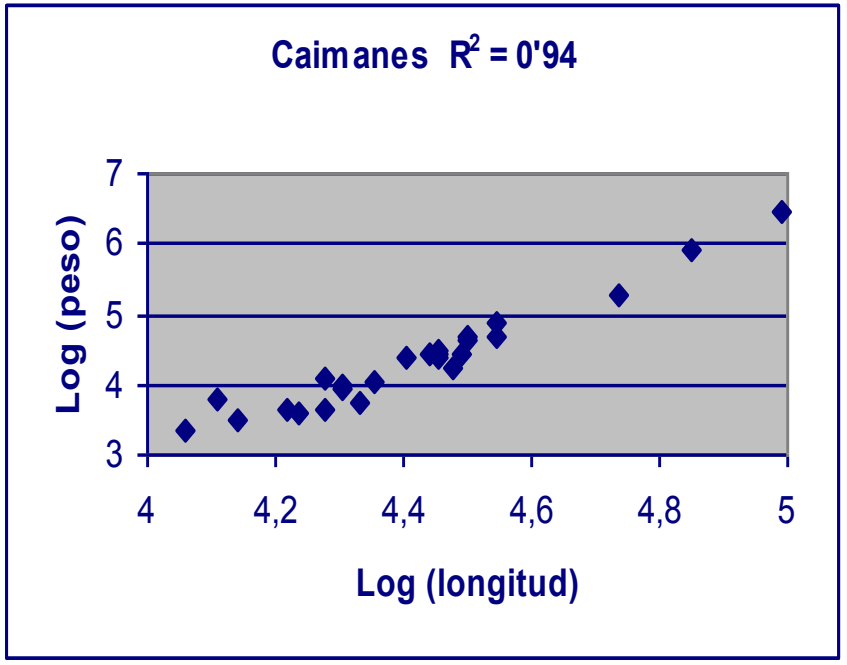
¿Qué función representa mejor el peso (Y) en función de la longitud (X)?

$$Y = \beta_0 + \beta_1 X^3$$

$$Y = k X^{\beta 1}$$

$$Y = ke^{\beta 1 X}$$





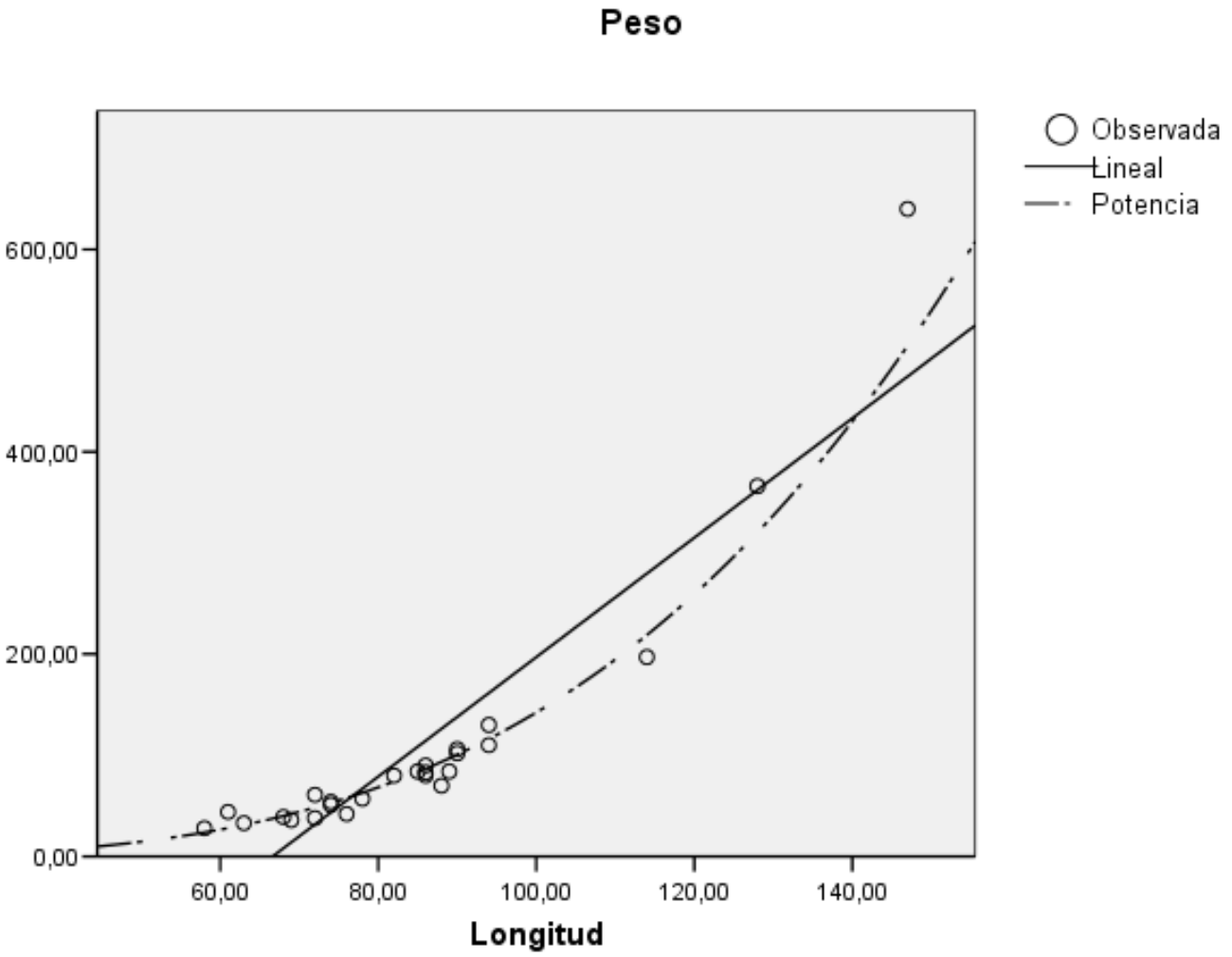
¿Qué modelo representa cada una de estas gráficas?

R^2 es la estimación del coeficiente de correlación al cuadrado

¿Qué efecto tendrán sobre el ajuste los 3 caimanes grandes?

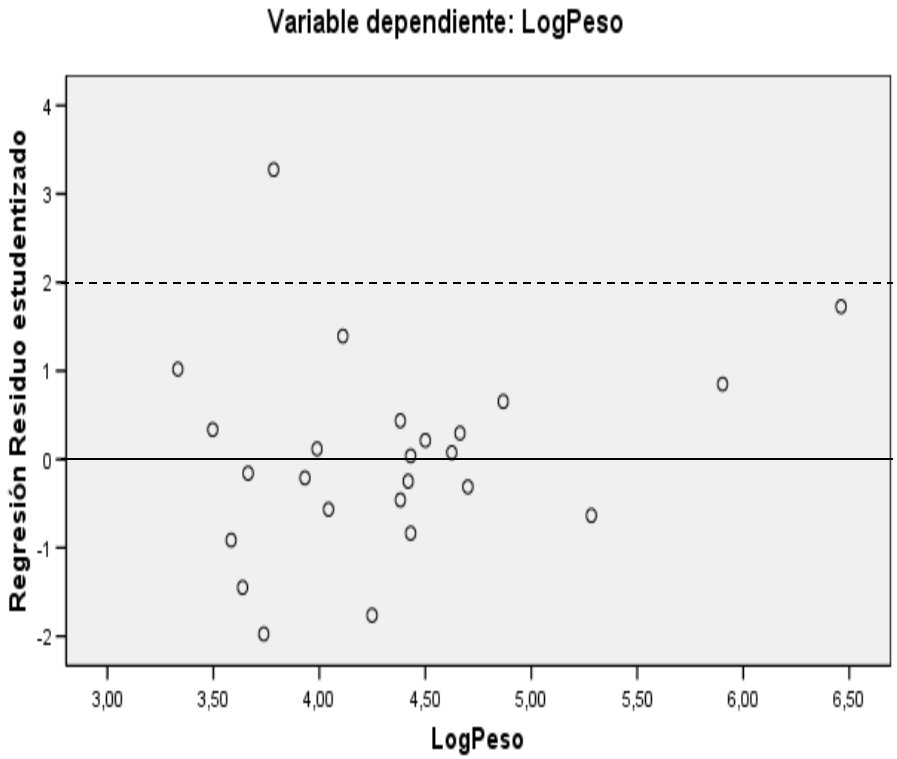
Modelo $\mathbf{Y = k X^{\beta 1}}$

Equivalente al ajuste lineal $\mathbf{Log(Y) = \beta_0 + \beta_1 Log(X)}$

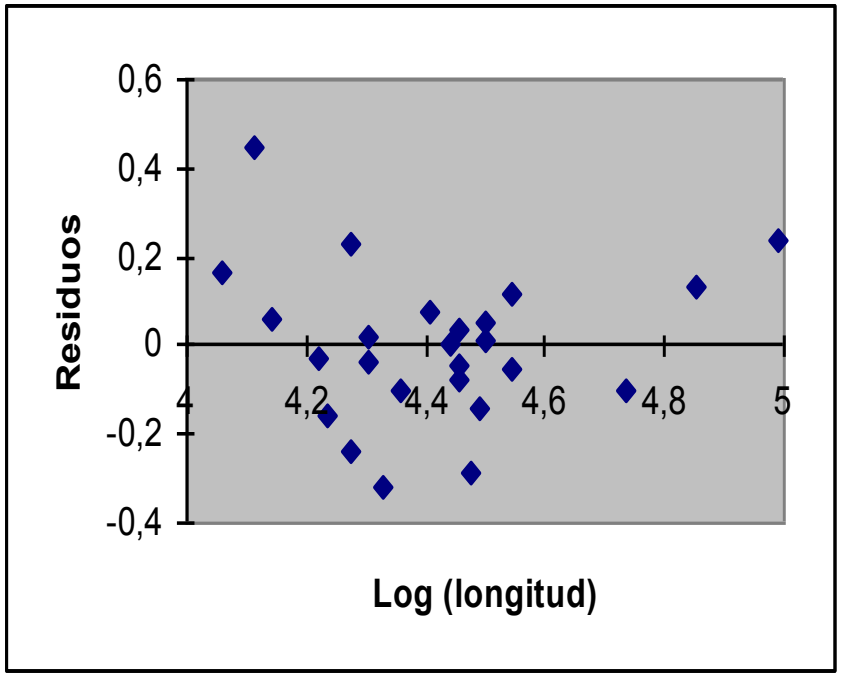


Modelo $Y = k X^{\beta 1}$: análisis de los residuos

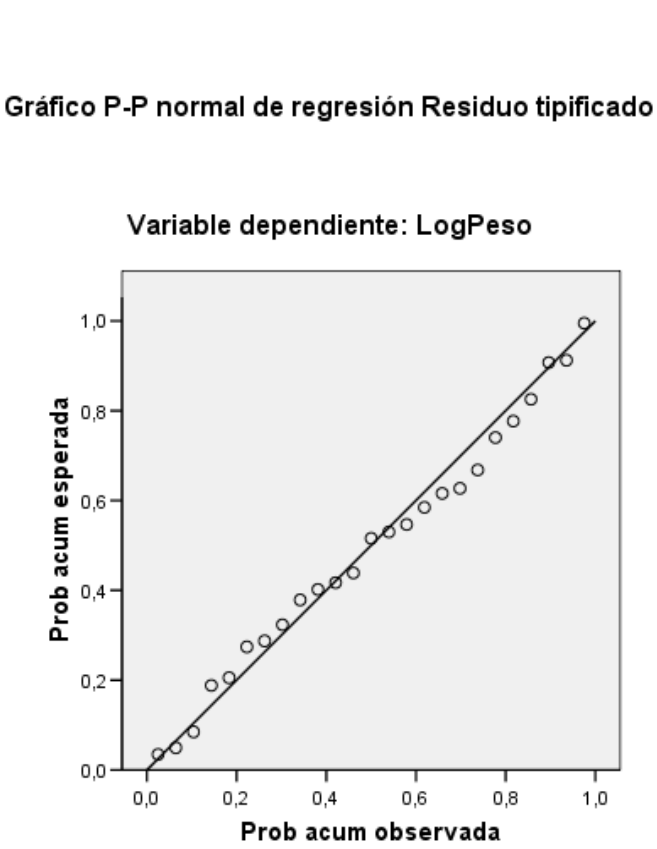
Residuos tipificados sobre Log(peso)

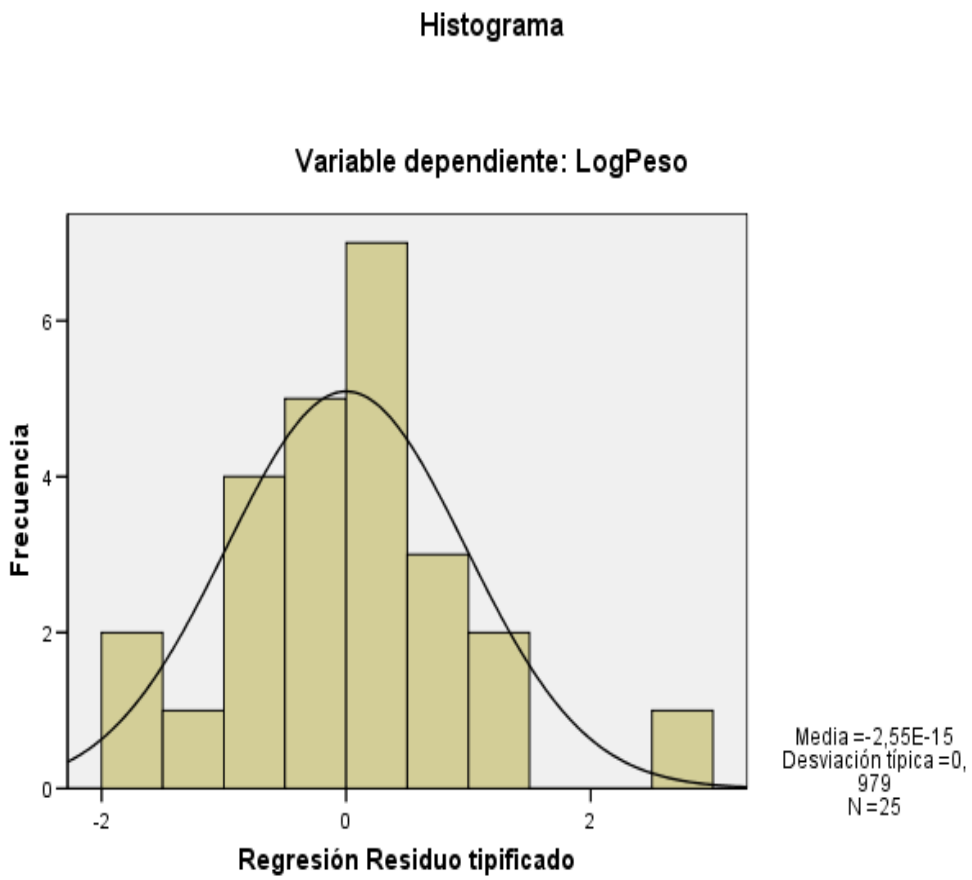


Residuos brutos sobre Log(longitud)



Modelo $Y = k X^{\beta 1}$: análisis de los residuos (Normalidad)





Modelo $Y = \beta_0 + \beta_1 X^3$: análisis de los residuos

Residuos tipificados

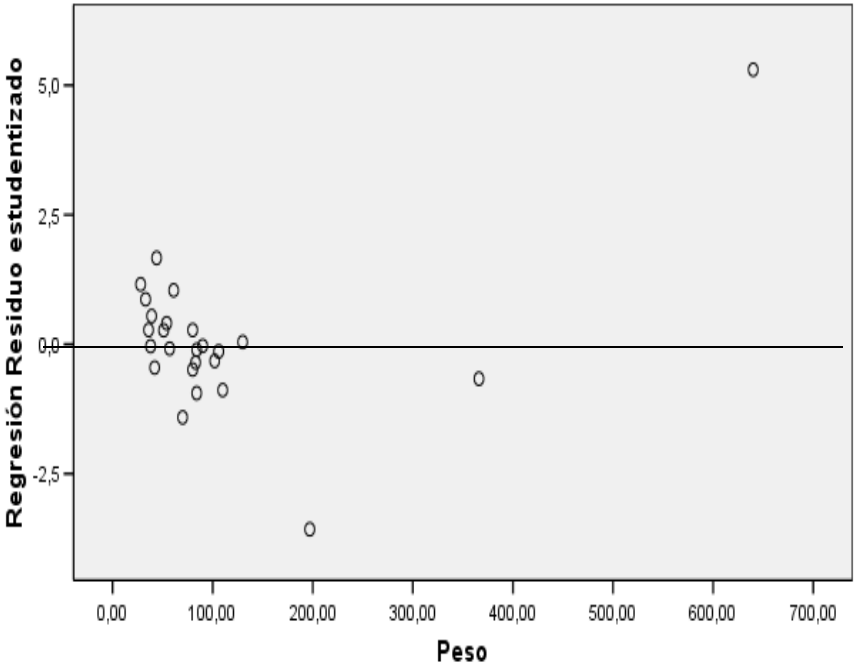
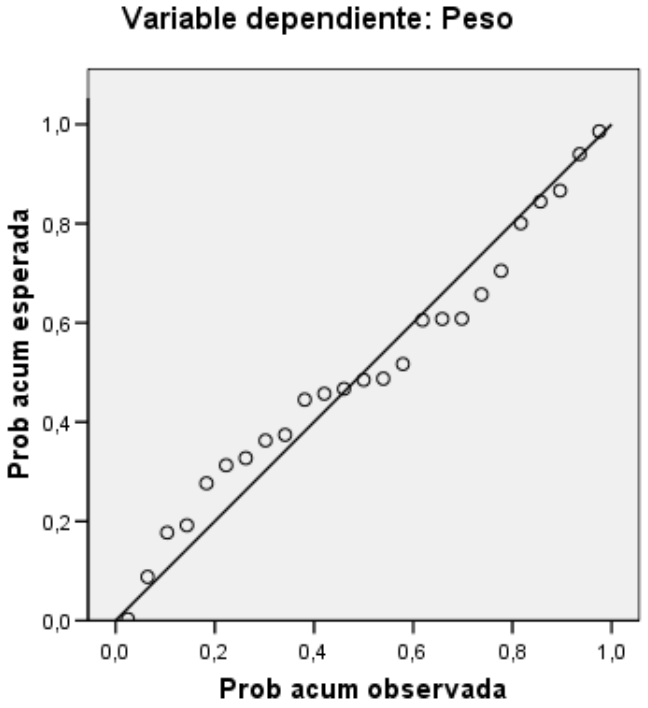


Gráfico P-P normal de regresión Residuo tipificado



CONTRASTES DE LA REGRESIÓN: t

$H_0 : \beta_1 = 0$ (Los valores de la X no influyen en los valores de Y en una relación lineal)

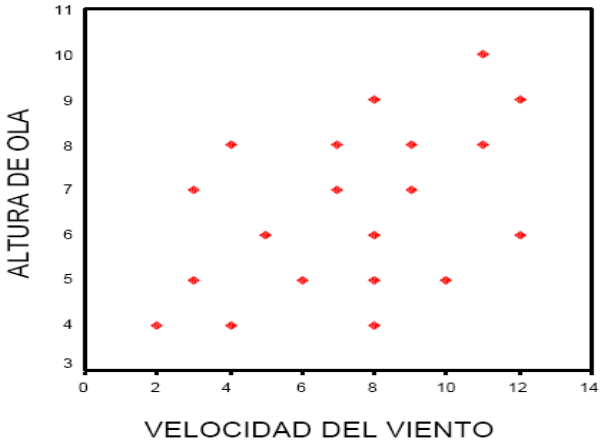
$H_1 : \beta_1 \neq 0$

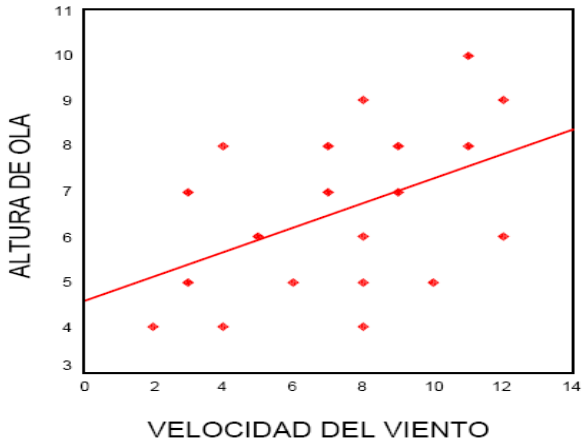
Con nivel de significación α rechazamos H_0 si el cero no está en el intervalo de confianza:

Error típico

$$IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{n-2;\alpha/2} S_R \sqrt{\frac{1}{n v_x}}\right)$$

Ejemplo 2. Altura de ola en función de la velocidad del viento





Coeficientes[a]

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)	4,549	,981		4,639	,000	2,489	6,609
	VELOCIDAD DEL VIENTO	,272	,124	,461	2,204	,041	,013	,532

a. Variable dependiente: ALTURA

Ejemplo 1. Caimanes con la transformación doble log

Modelo 1	*Coeficientes*	*Error típico*	*t*	*p-valor*	*Inferior 95%*	*Superior 95%*
Intercepción	-10,175	0,732	-13,907	1,1E-12	-11,688	-8,661
Log(Longitud)	3,286	0,165	19,868	5,59E-16	2,944	3,628

Curva de regresión estimada:

Log Y = -10,175 + 3,286 Log X

o equivalentemente:

$Y = e^{-10,175} X^{3,286} = 0,0000381 X^{3,286}$

CONTRASTES DE LA REGRESIÓN: ANOVA

Descomposición de la variabilidad en regresión

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \underbrace{e_i}_{y_i - \hat{y}_i}$$

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) \quad \text{(restando } \bar{y}\text{)}$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad \text{(elevando al cuadrado y sumando)}$$

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

SCT = SCE + SCR

SCE Suma de cuadrados explicada (variabilidad de y debida a su relación lineal con la x)

SCR Suma de cuadrados residual (variabilidad de y respecto a la recta ajustada)

SCT Suma de cuadrados total (variabilidad total de la y)

TABLA ANOVA

Suma de cuadrados	G.l.	Varianza	Estadístico	p-valor
$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$\frac{SCE}{1}$	$F = \frac{SCE/1}{SCR/(n-2)}$	¿?
$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-2$	$\frac{SCR}{n-2}$		
$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$			

H_0 : **El modelo de regresión lineal NO sirve para explicar la respuesta**

H_1 : **El modelo de regresión lineal SI sirve para explicar la respuesta**

A nivel de significación α, rechazamos cuando

$$F > F_{1,n-2,\alpha}$$

Coeficiente de determinación – R^2

Valoración de cuánto se ajustan los puntos a la recta

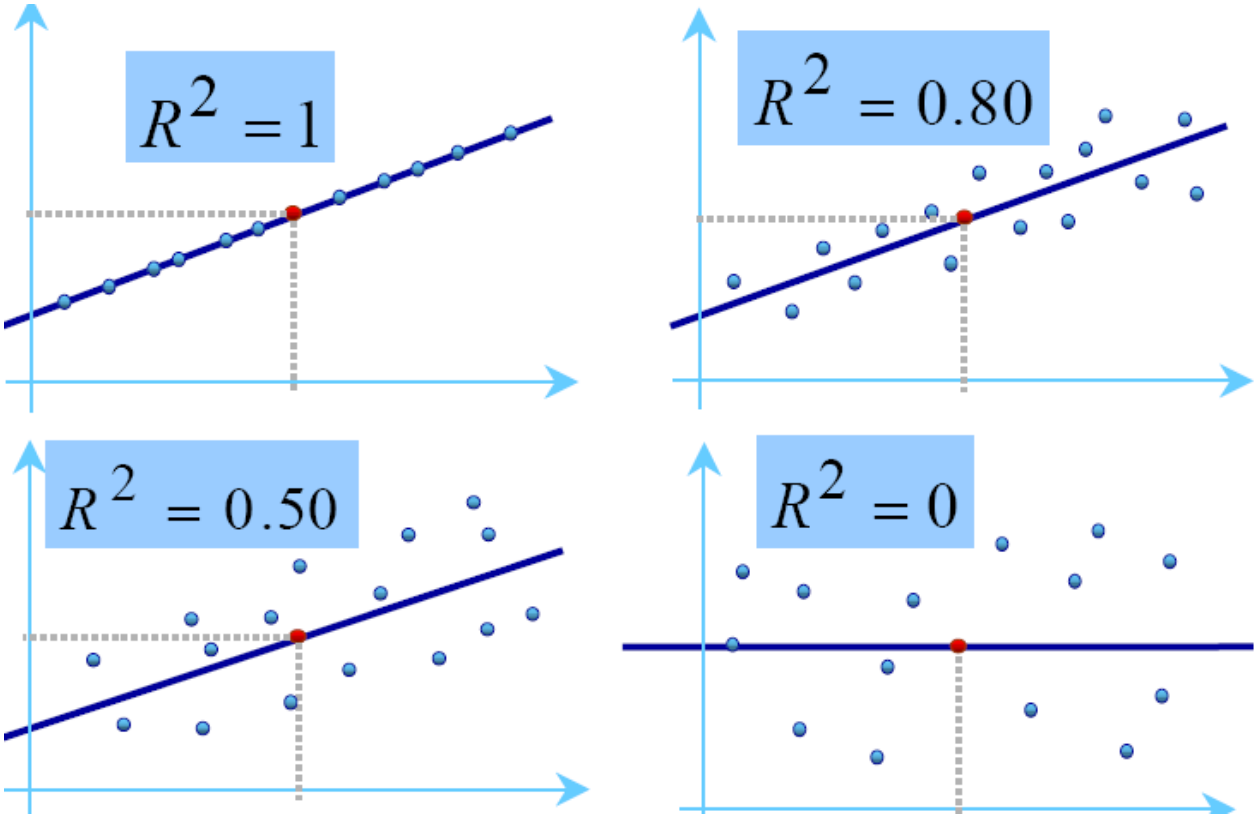
El **COEFICIENTE DE DETERMINACIÓN** es la proporción de variabilidad explicada por la regresión

R^2 = SCE /SCT

En REGRESIÓN SIMPLE el COEFICIENTE DE DETERMINACIÓN coincide con el COEFICIENTE DE CORRELACIÓN AL CUADRADO

R = $r = \frac{cov}{\sqrt{v_x v_y}}$; $SCR = nv_y(1 - r^2)$

Coeficiente de determinación – R^2



Comentarios:

- El contraste de la regresión supone que la relación (más o menos fuerte) es LINEAL. Por tanto, **si no rechazamos** la hipótesis nula lo único que podemos decir es que **no hemos encontrado evidencia de que exista una relación lineal**, puede existir una relación no lineal…

- En REGRESIÓN SIMPLE el contraste ANOVA coincide exactamente con el contraste de la *t* para el coeficiente de la variable regresora

Ejemplo 2. Altura de ola en función de la velocidad del viento

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,461[a]	,213	,169	1,65949

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	13,380	1	13,380	4,858	,041[a]
	Residual	49,570	18	2,754		
	Total	62,950	19			

a. Variables predictoras: (Constante). VELOCIDAD DEL VIENTO

b. Variable dependiente: ALTURA

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	4,549	,981		4,639	,000
	VELOCIDAD DEL VIE	,272	,124	,461	2,204	,041

Ejemplo 1. Caimanes con la transformación doble log

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,972[a]	,945	,943	,17531

a. Variables predictoras: (Constante), LogLongitud

b. Variable dependiente: LogPeso

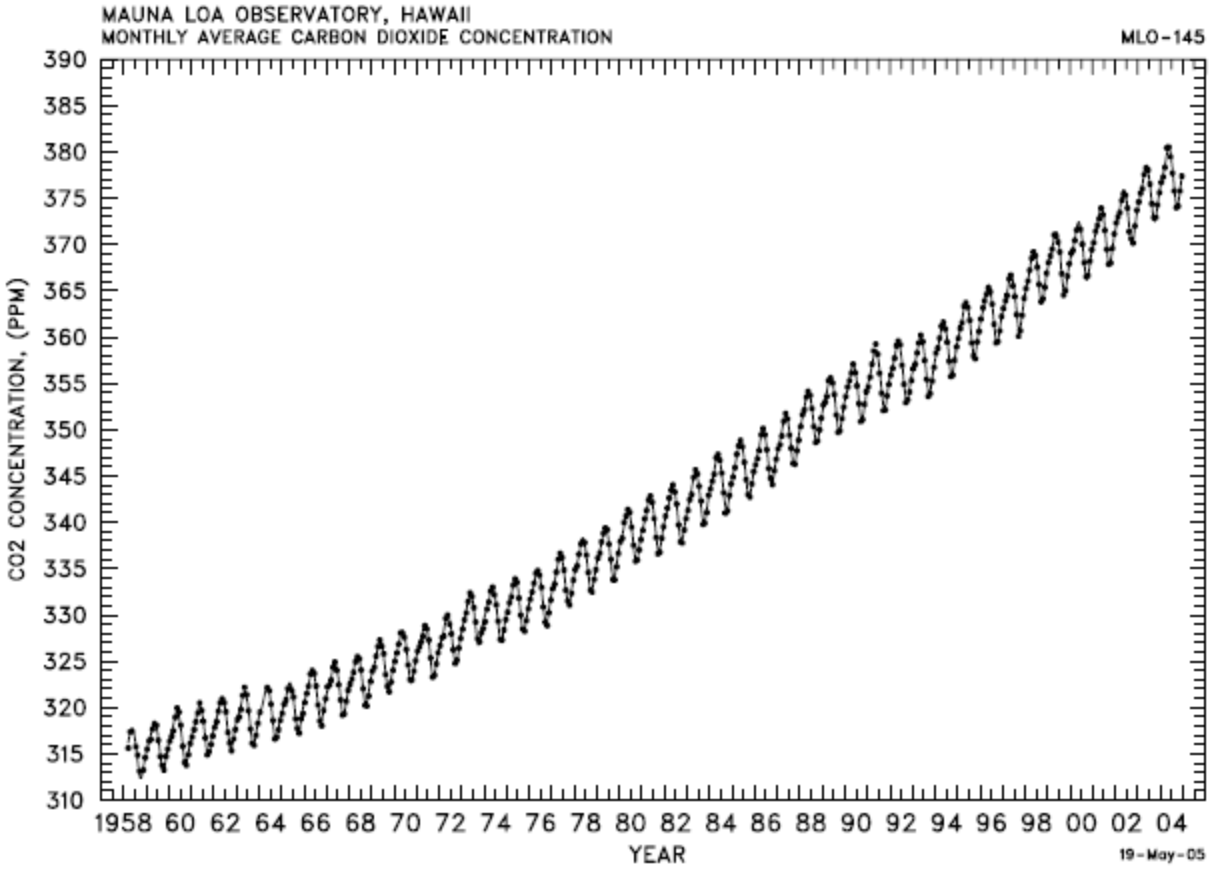
ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	12,132	1	12,132	394,729	,000[a]
	Residual	,707	23	,031		
	Total	12,838	24			

a. Variables predictoras: (Constante), LogLongitud

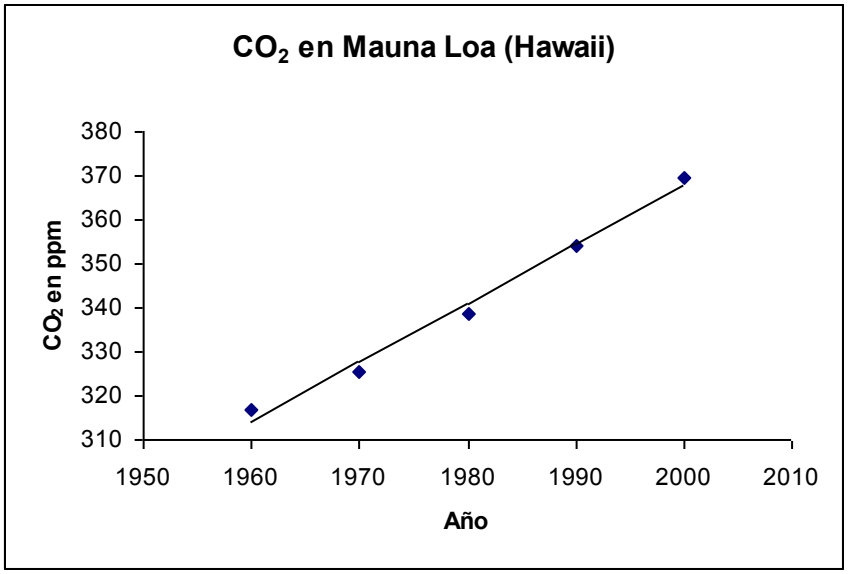
b. Variable dependiente: LogPeso

Ejemplo 3



Datos extraidos de: C. D. Keeling, T. P. Whorf & CO2 Research Groups (SIO); U. California, La Jolla; en http://cdiac.ornl.gov/trends/co2/sio-mlo.htm

Año	CO_2
1960	316,91
1970	325,68
1980	338,69
1990	354,19
2000	369,47



Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,9947
Coeficiente de determinación R^2	0,9894
R^2 ajustado	0,9858
Error típico	2,5288
Observaciones	5

ANÁLISIS DE VARIANZA						
	g. l.	*S. C.*	*M. C.*	*F*	*F-crit*	
Regresión	1	1785,70	1785,70	**279,23**	**0,0005**	
Residuos	3	19,19	**6,40**			
Total	4	1804,88				
	Coef.	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	**300,90**	2,65	113,45	0,0000	292,46	309,34
Variable X 1	**13,36**	0,80	16,71	**0,0005**	**10,82**	**15,91**

Diagnóstico de las hipótesis del modelo

Si las hipótesis del modelo son ciertas, entonces los residuos son aproximadamente

1. **Normales**
2. **Media cero**
3. **Independientes**
4. **Varianza constante**
5. **No hay residuos atípicos**

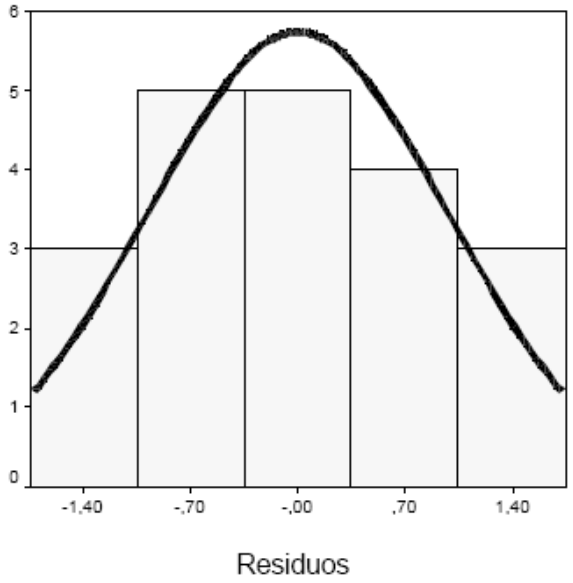
Podemos utilizar contrastes y gráficos para ver si hay EVIDENCIA CLARA en contra de alguna de las hipótesis

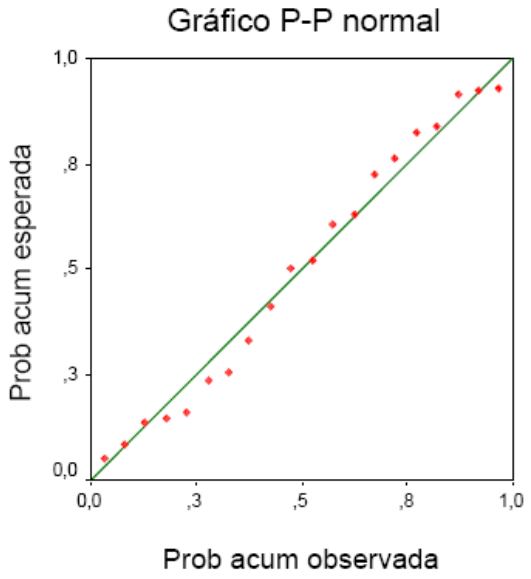
Normalidad
- Histograma de los *residuos tipificados*
- Q-Q plot de los *residuos tipificados*
- Test de K-S de los *residuos tipificados*

Linealidad / Homocedasticidad
- Diagrama de dispersión de los *residuos tipificados* frente a los *valores pronosticados/ajustados*

Tienen que estar entre -2 y 2, en una nube de puntos sin forma

Ejemplo 2. Altura de ola en función de la velocidad del viento





Prueba de Kolmogorov-Smirnov para una muestra

			Unstandardized Residual
N			20
Parámetros normales	a,b	Media	,0000000
		Desviación típica	1,61522698
Diferencias más extremas		Absoluta	,101
		Positiva	,101
		Negativa	-,082
Sig. asintót. (bilateral)			,987

Predicciones a partir del modelo ajustado

Una vez aceptado el modelo de regresión, podemos plantearnos realizar estimaciones y predicciones sobre distintas características de la Y dado un valor fijo de X que denominaremos x_0

Analizaremos dos opciones:

- Estimación de $E(Y/X=x_0)$ valor medio de Y para $X=x_0$
- Predicción de un valor de Y para $X=x_0$

En ambos casos la mejor **estimación puntual** es el valor de Y predicho por la recta de regresión ajustada:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

¿Dónde está la diferencia?

Ejemplo: para una misma velocidad del viento x_0 las olas podrán tener distintas alturas: recordemos que hemos aceptado una $N(\beta_0+\beta_1 x_0, \sigma)$

Estimación de la media de Y dado $X=x_0$
Estimación de la altura media que tendrán **todas** las olas para una velocidad del viento fija x_0

Predicción de un valor de Y dado $X=x_0$
Predicción de la altura de **una** ola para una velocidad del viento fija x_0

La estimación de la media será más precisa ya que compensamos la variabilidad de la Y para $X=x_0$

En la predicción de un único valor, a la variabilidad estadística se suma la variabilidad de los valores de la Y para $X=x_0$

Intervalos de confianza para la estimación y la predicción

Estimación de la media de Y dado $X=x_0$

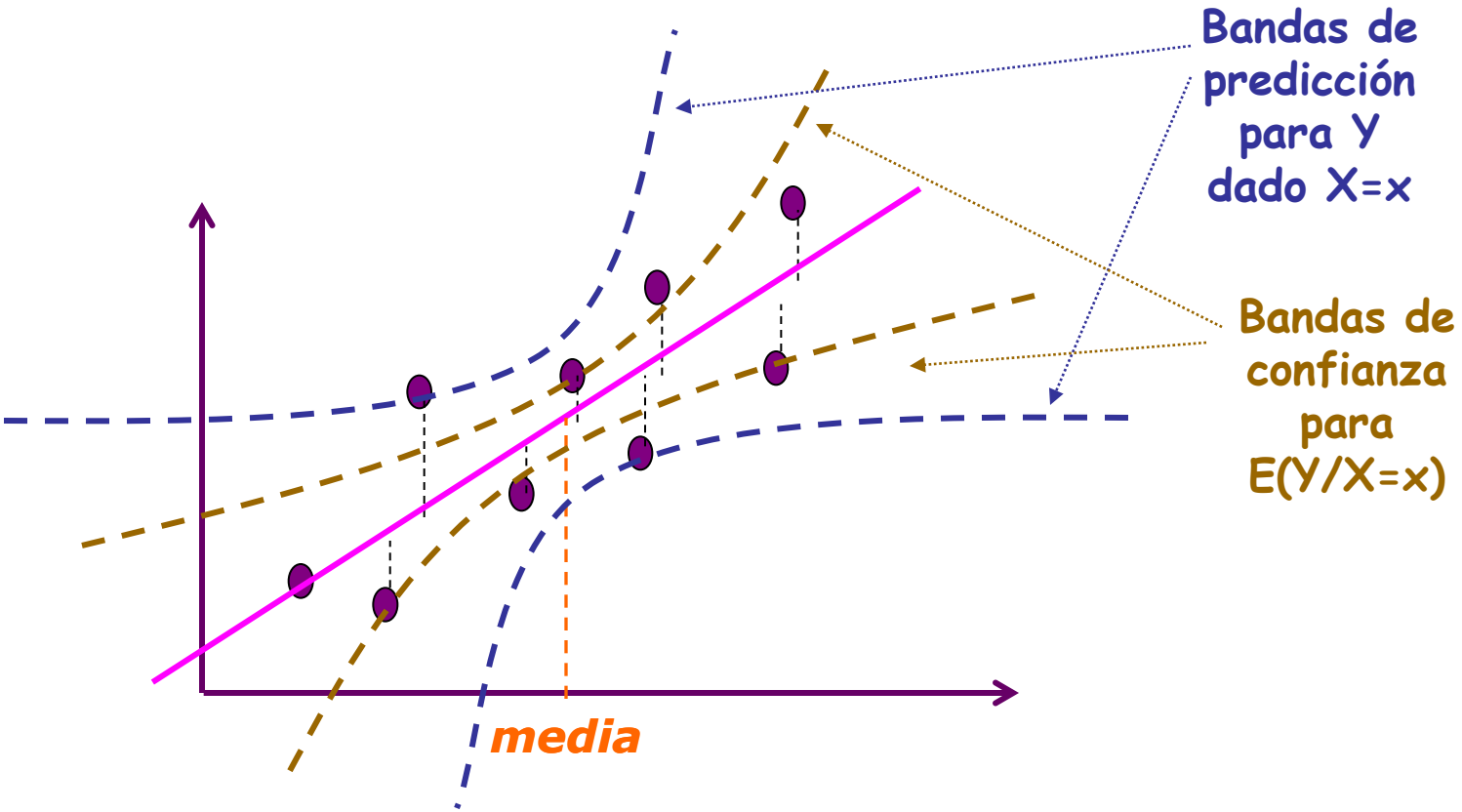
$$IC_{1-\alpha}(\text{estimación}) = \left(\hat{y}_0 \pm t_{n-2;\alpha/2} S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n v_x}} \right)$$

Predicción de un valor de Y dado $X=x_0$

Error típico

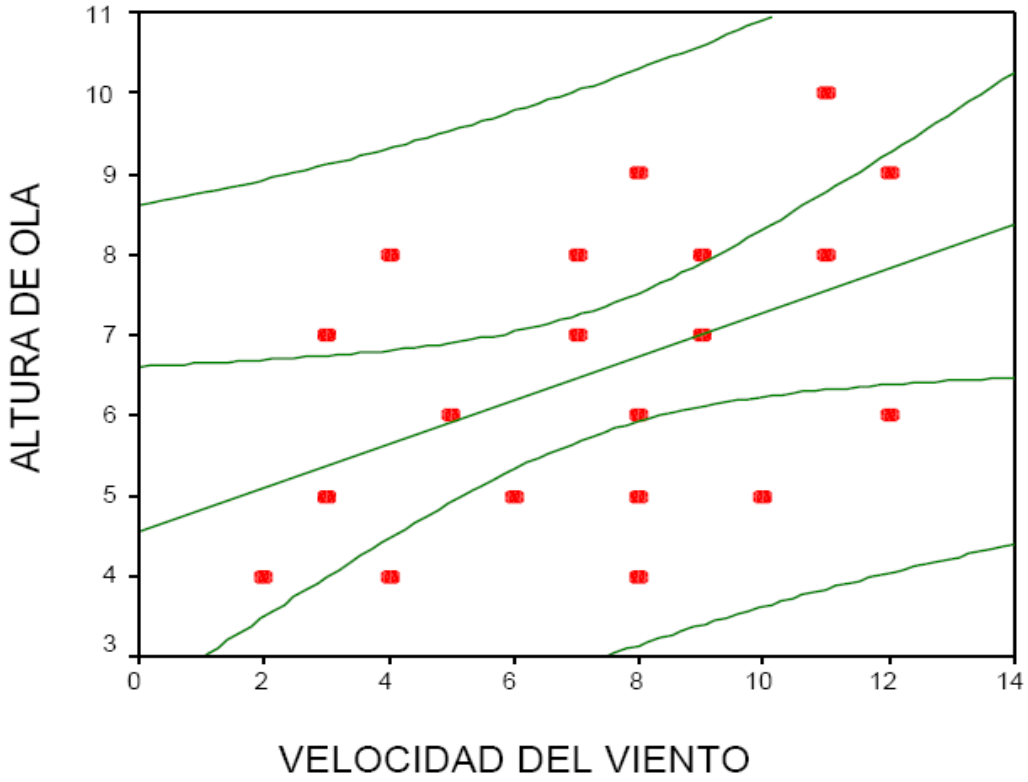
$$IC_{1-\alpha}(\text{predicción}) = \left(\hat{y}_0 \pm t_{n-2;\alpha/2} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n v_x}} \right)$$

Gráficamente: Bandas de confianza y de predicción



- Los dos bandas tienen la misma forma, siempre más estrechas en la media de las *x* donde hay más información

Ejemplo 2. Altura de ola en función de la velocidad del viento



Ejemplo 1. Caimanes con la transformación doble log

Curva de regresión estimada:

Log Y = -10,175 + 3,286 Log X

o equivalentemente:

$Y = e^{-10,175} X^{3,286} = 0,0000381 X^{3,286}$

¿Qué peso estimaríamos en media para los caimanes cuya longitud sea 100 pulgadas?

Respuesta: log (y_{100}) = 4,958 luego y_{100} = 142,25 libras

¿Que incremento del peso estimamos que resultaría de un incremento del 1% en la longitud?

log ($y_{1,01x}$) – log (y_x) = log ($y_{1,01x}/y_x$) = 3,286 log (1,01) = 0,0327

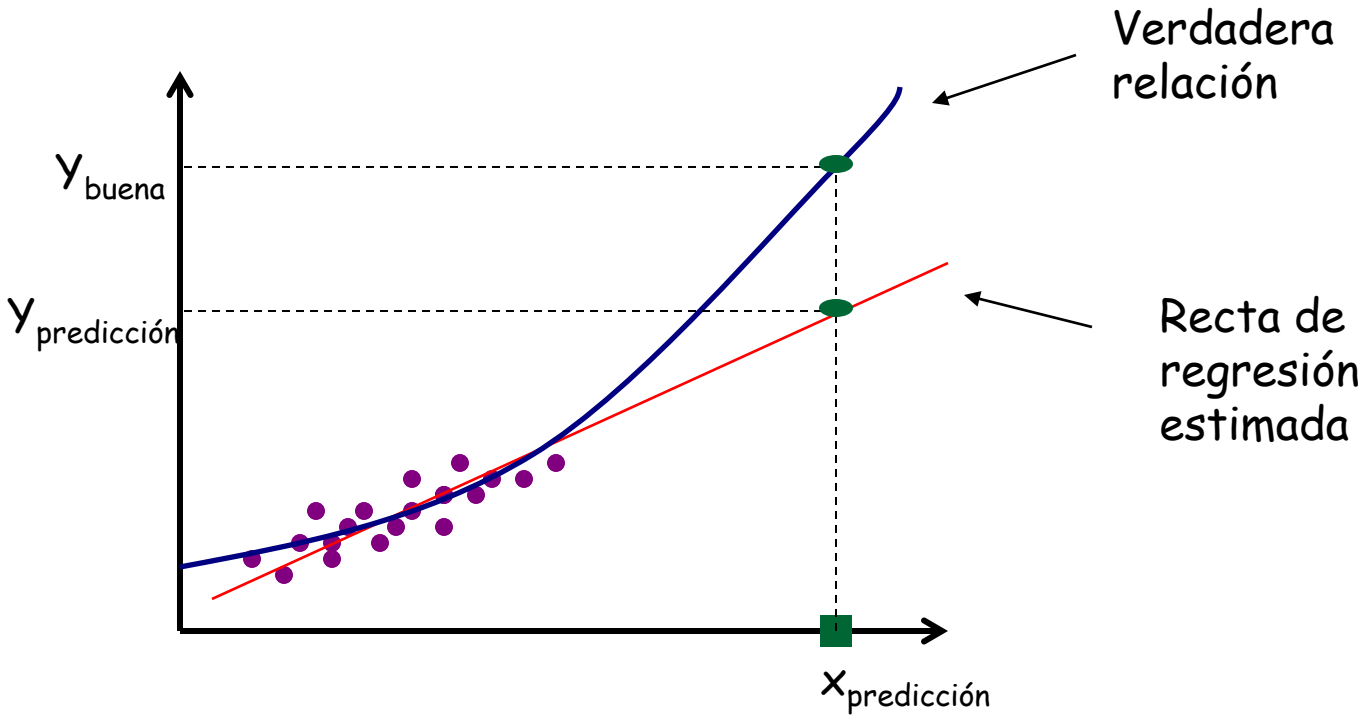
luego $y_{1,01x} = y_x\, e^{0,0327} = y_x\, 1,0332$ el peso se incrementaría en un 3,32%

Algunos abusos que se pueden cometer en la regresión

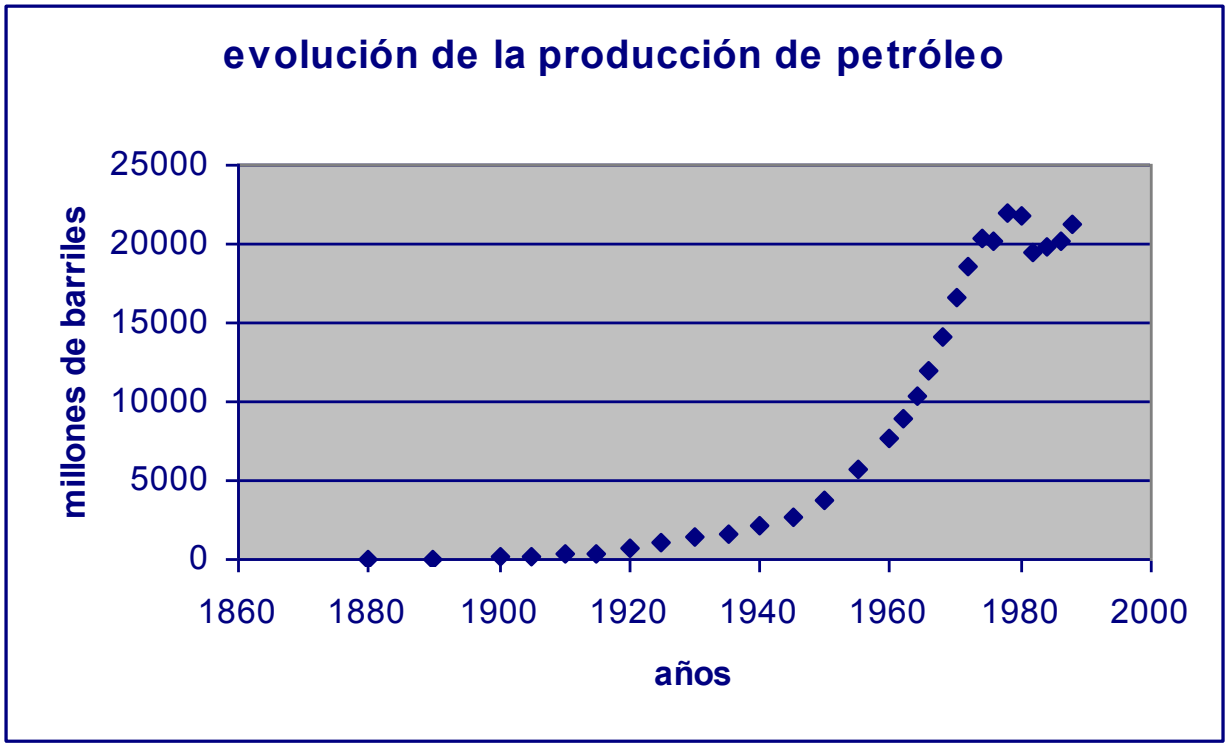
- Extrapolación
- Generalización
- Correlación ecológica
- Causalidad

Extrapolación

Aplicar el modelo a valores de la variable explicativa alejados de los observados

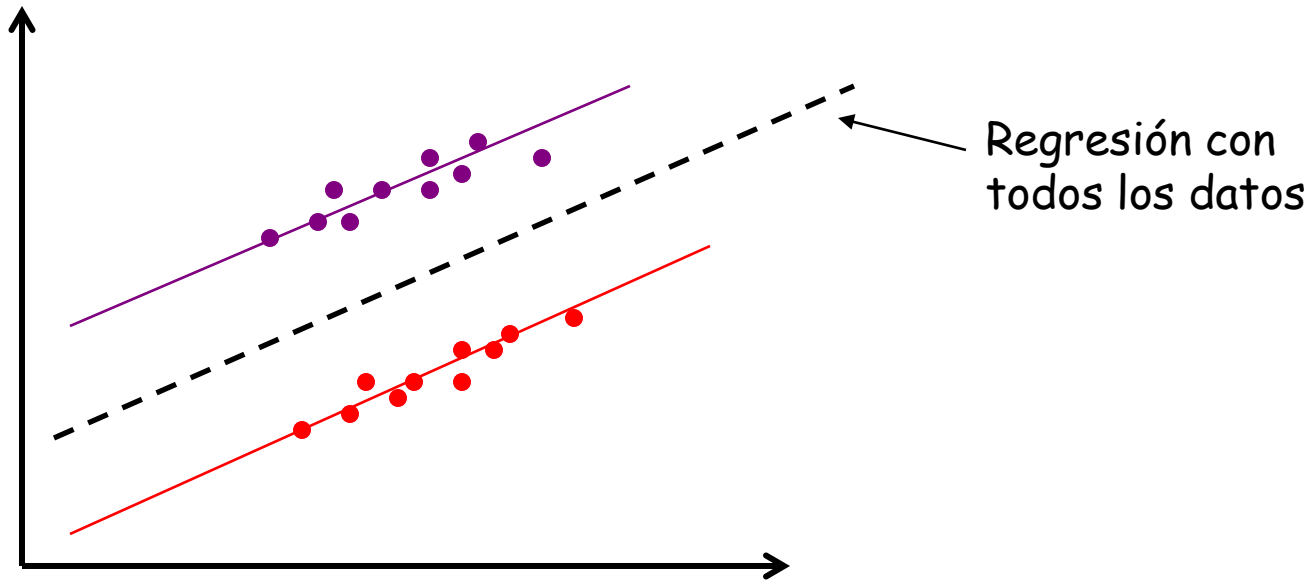


Ejemplo. Evolución de la producción de petróleo

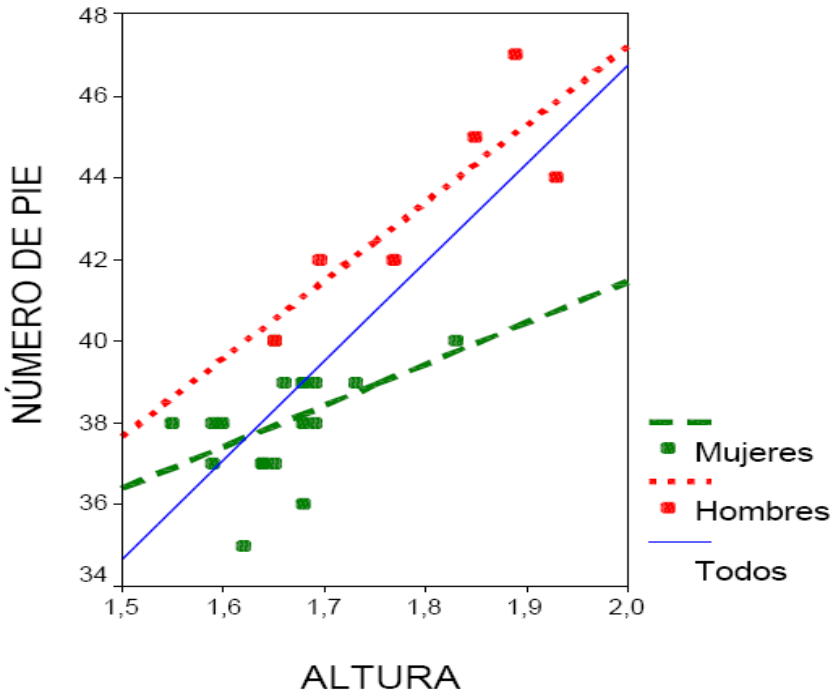


Generalización

Utilizar un único modelo para conjuntos de datos que proceden de distintas poblaciones

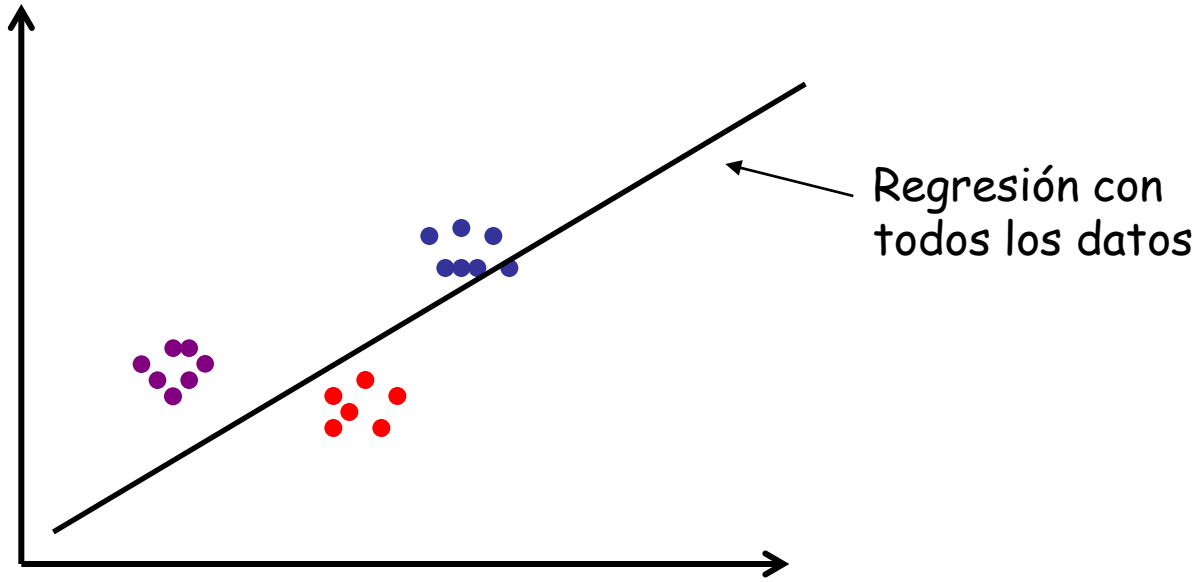


Ejemplo. Datos del número de pie en función de la altura de varios estudiantes de ambos sexos



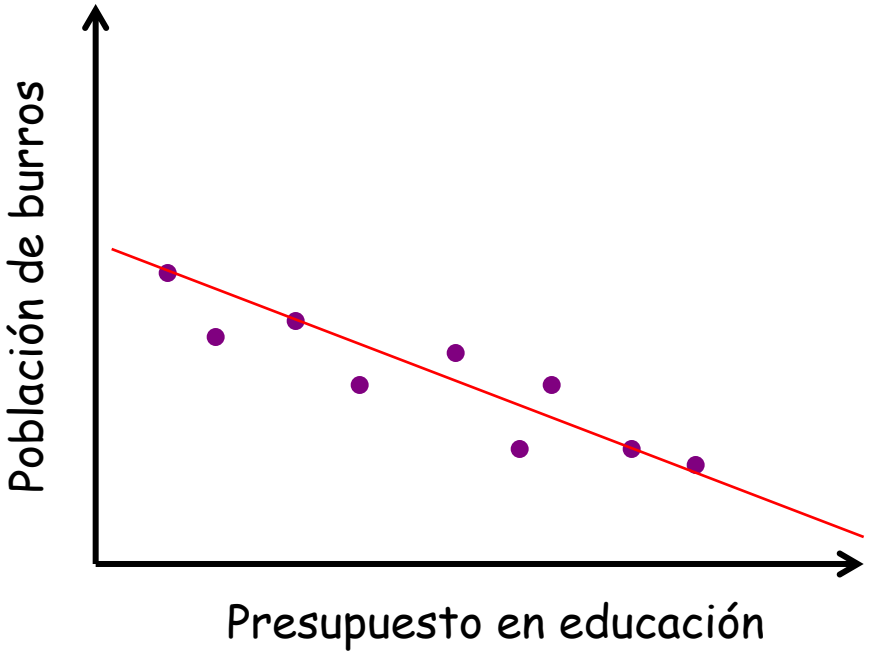
Correlación ecológica

Cuando no existe relación entre dos variables en ninguna de las poblaciones pero al juntar varias poblaciones aparece una falsa correlación



Causalidad

Admitir que existe una relación de causalidad entre las x's y las y's porque se ajusta bien un modelo



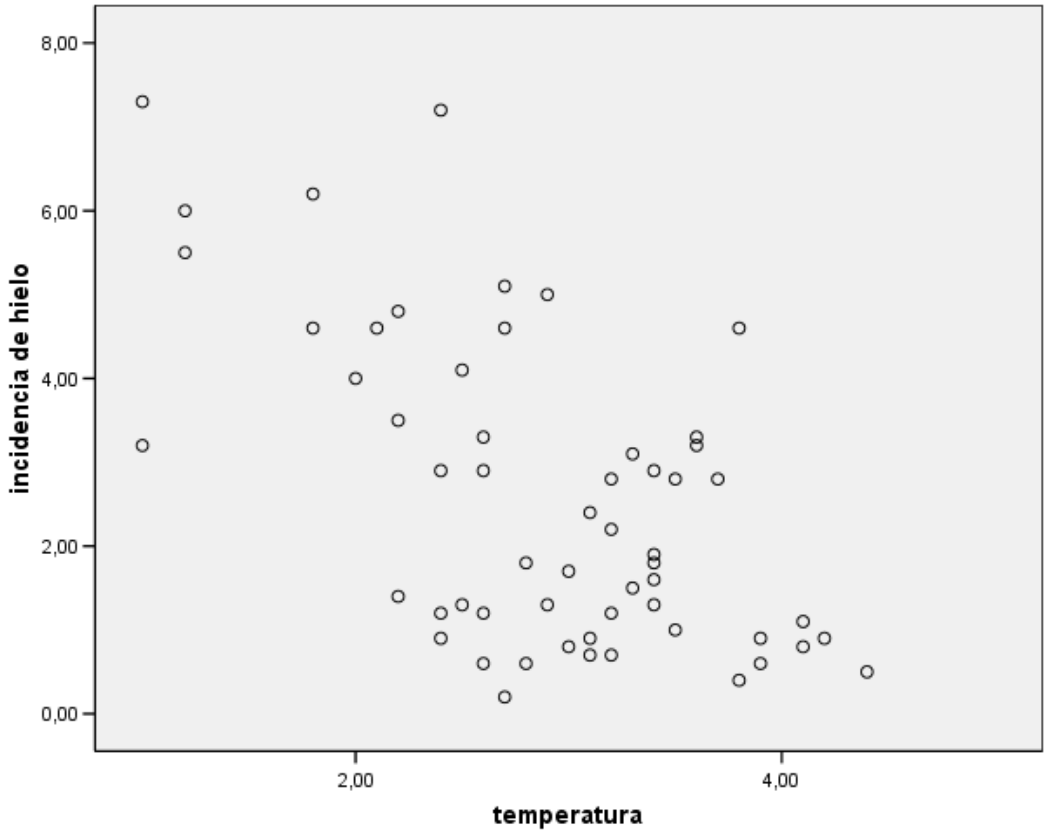
Correlación no implica Causalidad

Metodología para el análisis de la regresión

1. Identificar las variables respuesta y explicativa
2. "Comprobar" si son ciertas las hipótesis de *linealidad* y *homocedasticidad*
   - Diagrama de dispersión de los datos
   - Transformaciones de los datos
3. Estimar los parámetros del modelo
4. Hacer el contraste de la regresión:
   $H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$
   H_0 : **No** existe **relación lineal** entre Y y X
   H_0 : El **modelo no sirve** para explicar la respuesta
5. Diagnosis del modelo con los residuos:
   ¿Se cumple la hipótesis de *normalidad*?
6. ¿Hay alguna otra variable explicativa que pueda ser relevante y que podamos medir en los individuos de la muestra? SI $\longrightarrow$ 7. Regresión múltiple
7. Hacer predicciones con el modelo de regresión simple

Ejemplo 4.

Y = Incidencia de hielo (en meses por año) en las costas de Islandia en función de X = temperatura media anual.



n = 57 años

Ajuste del modelo lineal

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
temperatura	57	1,00	4,40	2,8947	,79066
incidencia de hielo	57	,20	7,30	2,5561	1,84556
N válido (según lista)	57				

Coeficientes[a]

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)	6,573	,759		8,661	,000	5,052	8,094
	temperatura	-1,388	,253	-,595	-5,484	,000	-1,895	-,881

a. Variable dependiente: incidencia de hielo

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	67,422	1	67,422	30,070	,000[a]
	Residual	123,319	55	2,242		
	Total	190,740	56			

a. Variables predictoras: (Constante), temperatura

b. Variable dependiente: incidencia de hielo

Normalidad

Histograma

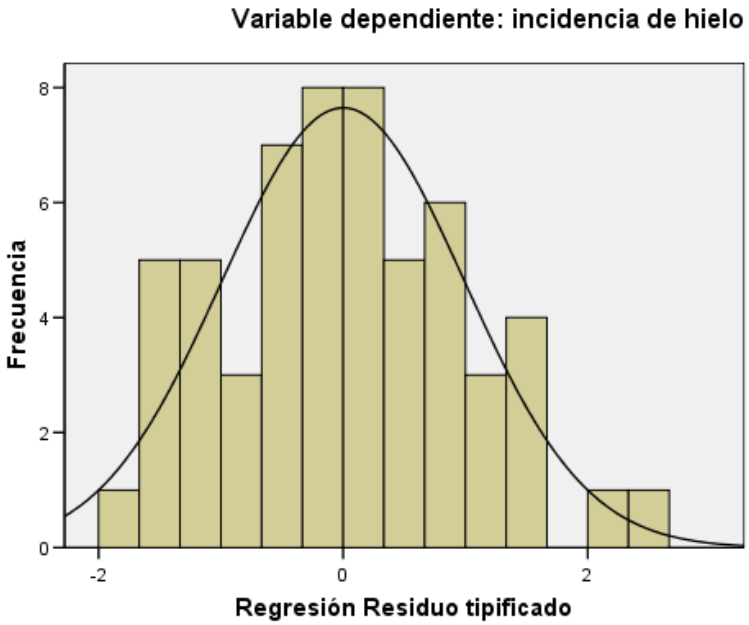
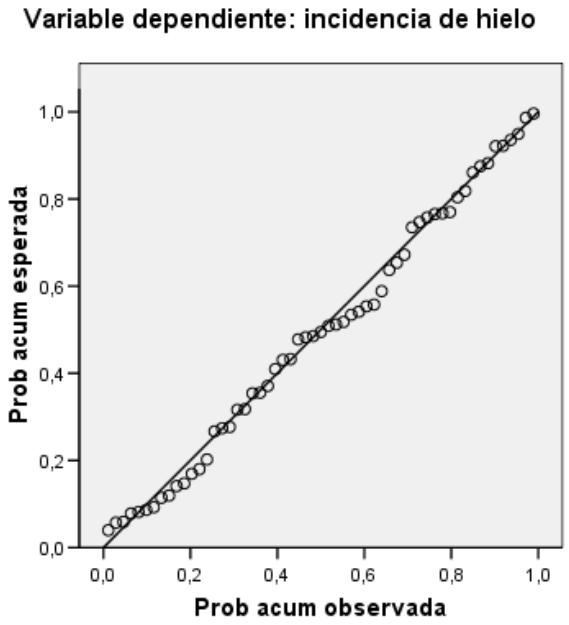
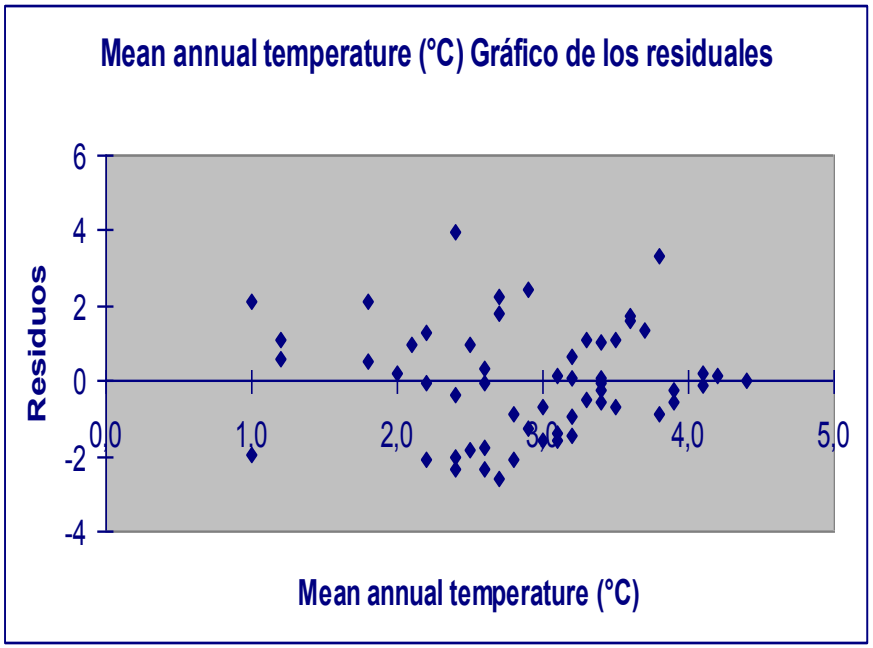
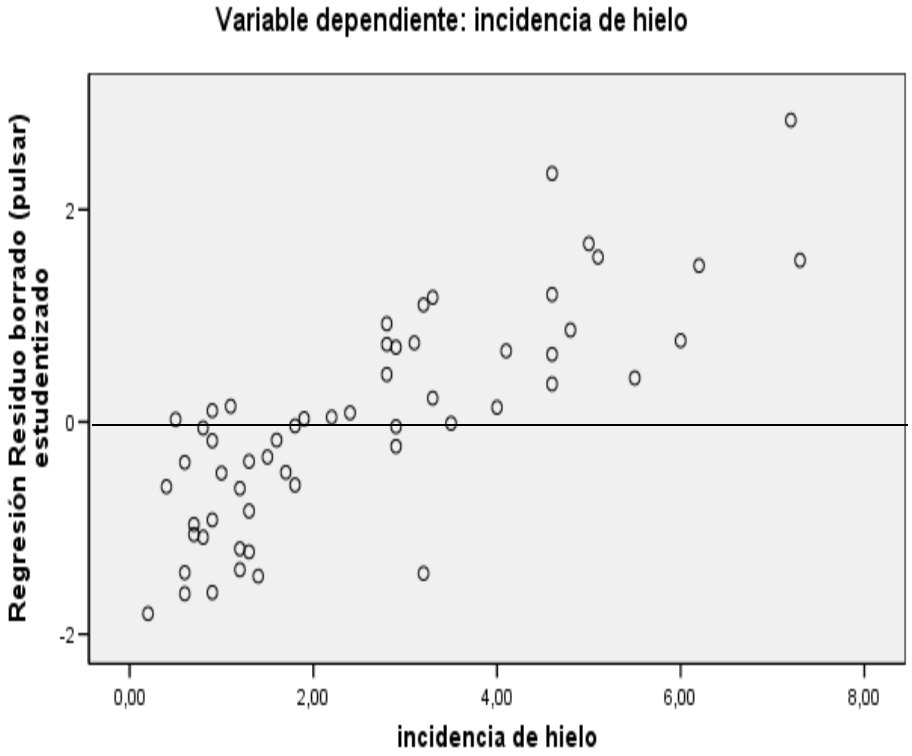


Gráfico P-P normal de regresión Residuo tipificado



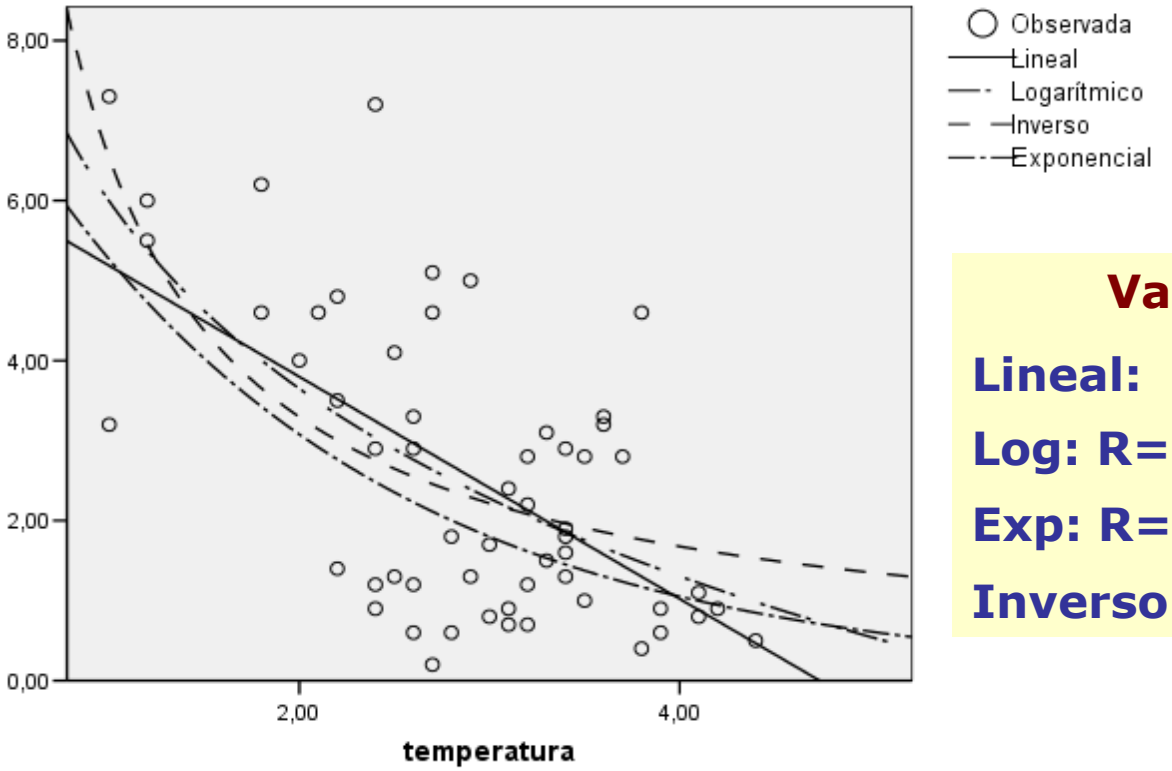
Dos gráficos de los residuos





¿Comentarios?

Otros modelos: transformaciones



Valores de R y F:

Lineal: R=0,595 F=30,07

Log: R= 0,609 F= 32,384

Exp: R= 0,514 F = 19,7

Inverso: R=0,586 F=28,7

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típico	Beta	t	Sig.
ln(temperatura)	-3,382	,594	-,609	-5,691	,000
(Constante)	5,993	,635		9,440	,000

Predicciones

¿Qué incidencia de hielo esperamos de un año en que la temperatura global sea de 1ºC?

Respuesta con el **modelo lineal**: 6,573 – 1,388 = 5,185 meses al año
Intervalo de confianza 0.95 para la incidencia **media** de hielo:
5,185 -+ $t_{55,0.025}$ 1,497 (0,515) = 5,185 -+ 1,03 = (4,155, 6,215)

¿Qué efecto tendrá sobre la incidencia del hielo un incremento de un 1ºC en la temperatura?
Respuesta: la incidencia de hielo descenderá en 1,388 meses

Respuesta con el **modelo logarítmico**: 5,993 – 3,382 log(1) = 5,993 meses

¿Qué efecto tendrá sobre la incidencia del hielo el multiplicar la temperatura por 2?
Respuesta: la incidencia de hielo descenderá en 2,344 meses

y_x= incidencia de hielo pronosticada a temperatura x = 5,993-3,382 log(x)
y_{2x}= incidencia de hielo pronosticada a temperatura 2x = 5,993-3,382 log(2x)
y_x-y_{2x} = 3,382 log(2) = 2,344