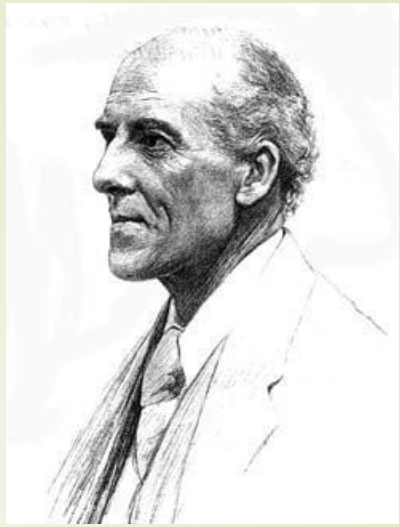
Tema 3
Regresión lineal simple

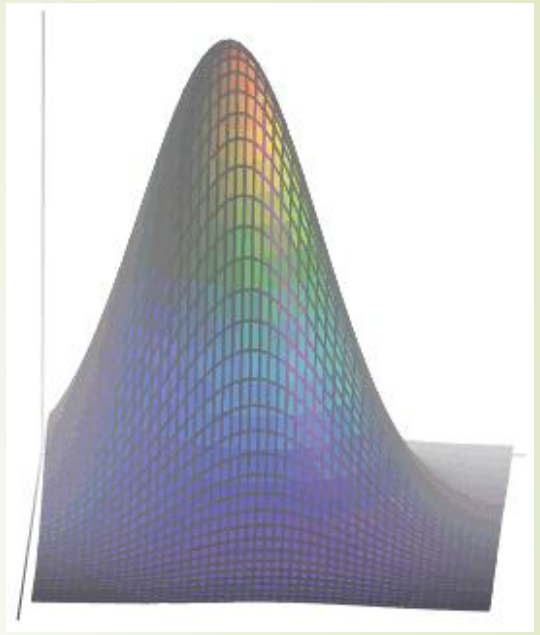


- Planteamiento del modelo de regresión lineal simple.
- Estimación de los parámetros.
- Diagnosis de las hipótesis del modelo a través de los residuos.
- Extensión a otros modelos de regresión simple: modelos linealizables.
- Tabla ANOVA. Evaluación del ajuste: coeficiente de correlación y coeficiente de determinación.
- Estimación de valores esperados y predicción de nuevas respuestas.
- Utilización del SPSS.

Una explicación para mi vida, se debe a una combinación de dos características que he heredado: capacidad para trabajar mucho y capacidad para relacionar las observaciones de los demás

Karl Pearson1857-1936

REGRESIÓN LINEAL SIMPLE
Nuevos elementos

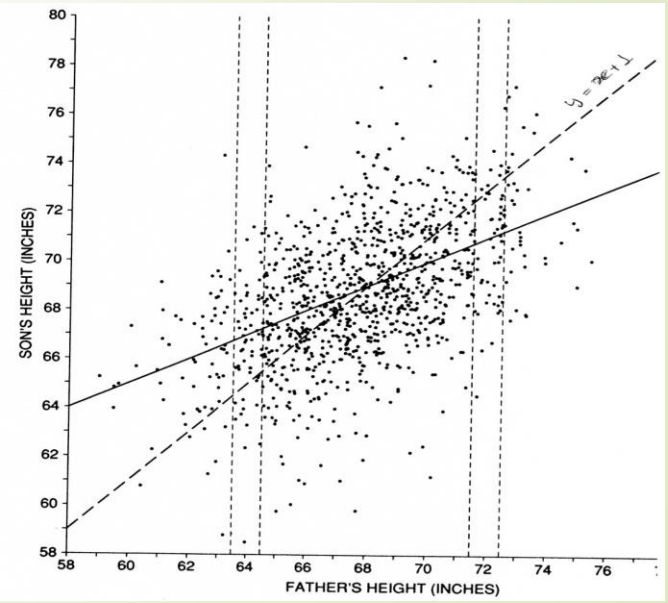


- La Normal bivariante

(Modelo de probabilidad conjunto para 2 variables)

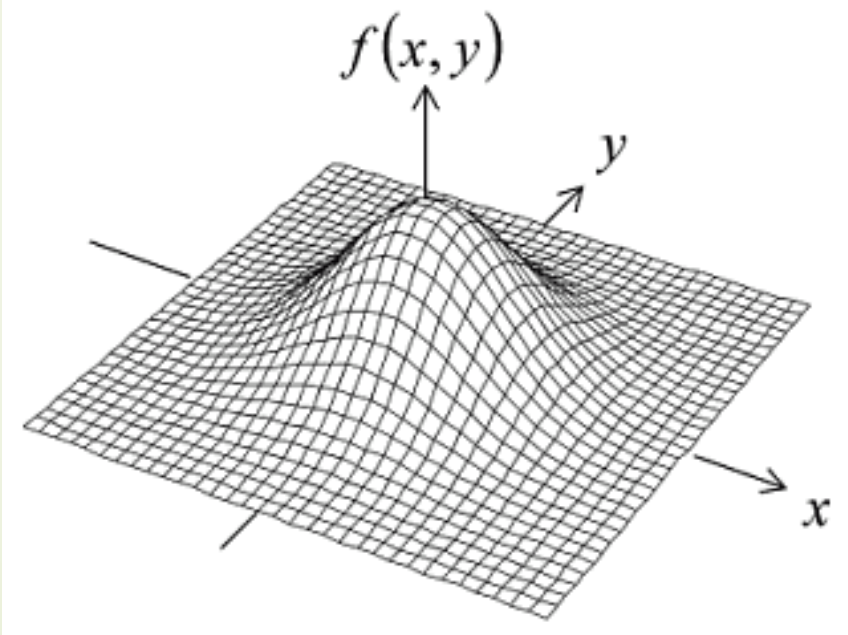
- Ajuste de una recta a una nube de puntos

(Análisis de datos de una muestra de 2 variables dependientes)



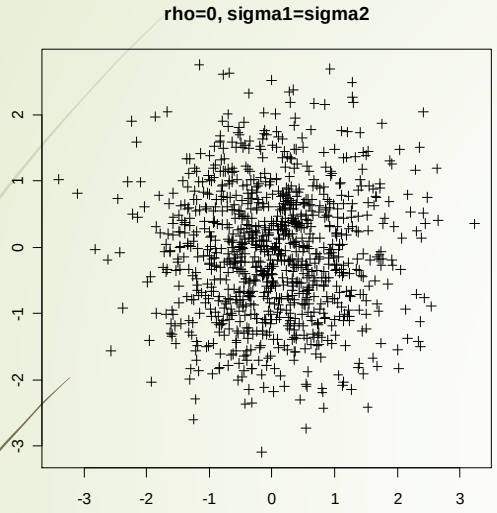
Distribución Normal Bivariante (X,Y) (parámetros $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$)

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

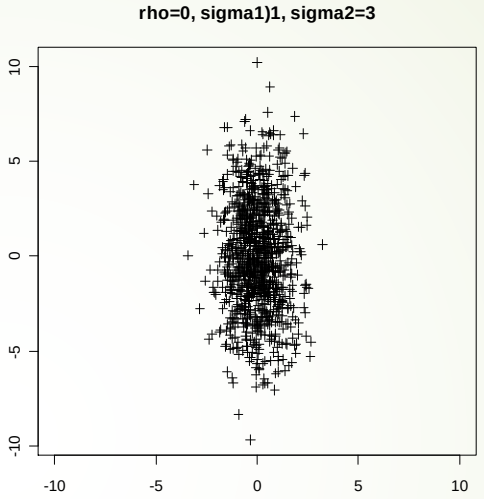


μ_1= E(X) μ_2 = E(Y) σ_1^2 = Var(X) σ_2^2 = Var(Y) ρ = **Coef. Correlación** (X,Y)

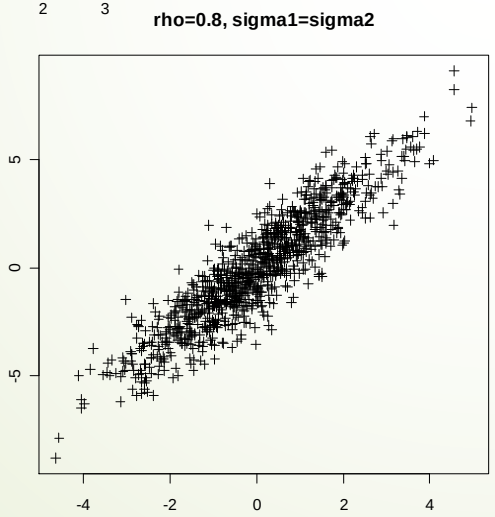
Distribución Normal Bivariante (simulación de datos)



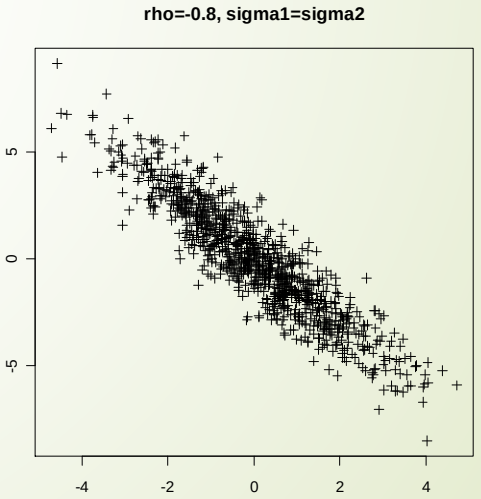
$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = 0$



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



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



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

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

La variable explicativa (X)
La variable dependiente a explicar (Y)

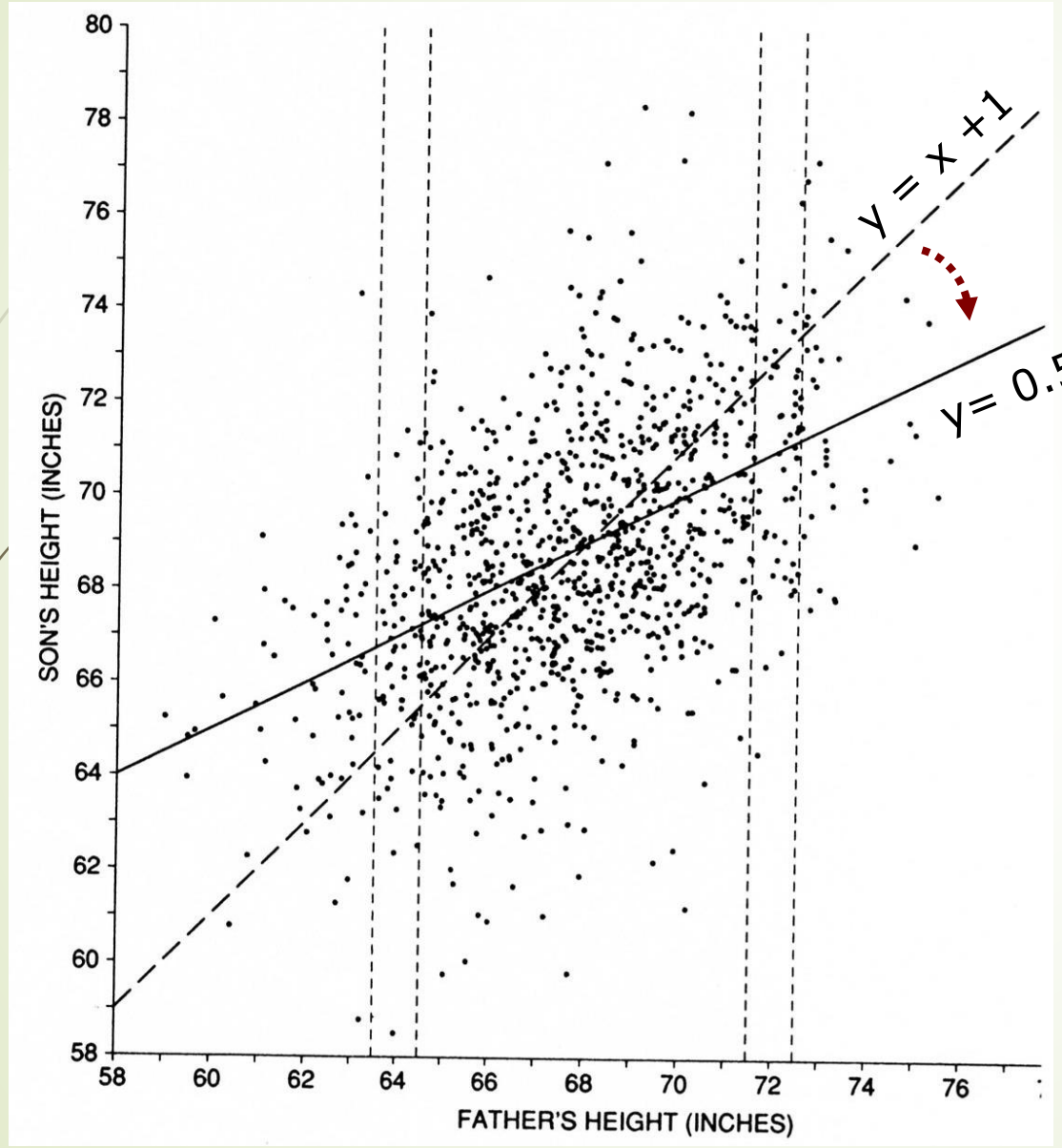
Y tratan de explicar la Y mediante una función lineal de los valores de la X representada por la recta

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

Para ello dispondremos:

De un modelo de probabilidad y de n pares de datos (x_i, y_i) que suponemos que provienen del modelo establecido y que se representan como una nube de puntos

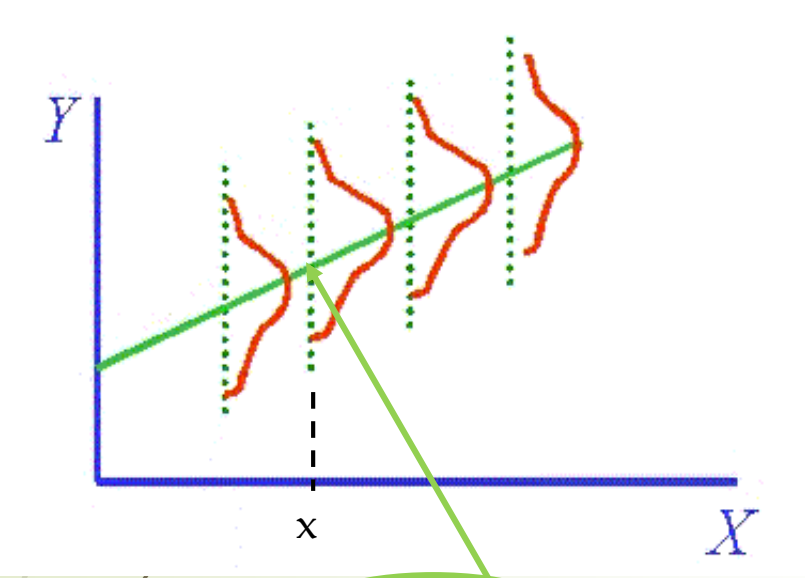
El origen:On the laws of inheritance in man
Karl Pearson
Biometrika 1903



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 para la Regresión lineal simple

$$Y_x = \beta_0 + \beta_1 x + U_x$$



$Y_x = \beta_0 + \beta_1 x + U_x$

- Y_x es la variable aleatoria que representa los valores que obtendremos cuando X tome un valor x (altura de los posibles hijos de un padre que mide x)
- $\beta_0 + \beta_1 x$ es el valor esperado (medio) de Y cuando X = x , es decir la $E(Y_x)$ (altura media de los posibles hijos de un padre que mide x)
- U_x representa la variabilidad aleatoria de Y cuando X = x (no todos los hijos de un padre que mide x miden lo mismo)

Supondremos que $\mathbf{U_x}$ sigue una distribución $\mathbf{N(0, \sigma)}$ igual sea cual sea el valor de x; es decir tiene media 0 y desviación típica σ independiente del valor de x.

Parámetros del modelo de regresión simple

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

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

σ^2 Representa la variabilidad respecto a la recta

Estos parámetros están relacionados con los de la Normal bivariante (μ_1, μ_2, σ_1, σ_2, ρ)

Muestra (se obtendrán n parejas de observaciones)

Dos formas de obtener los datos

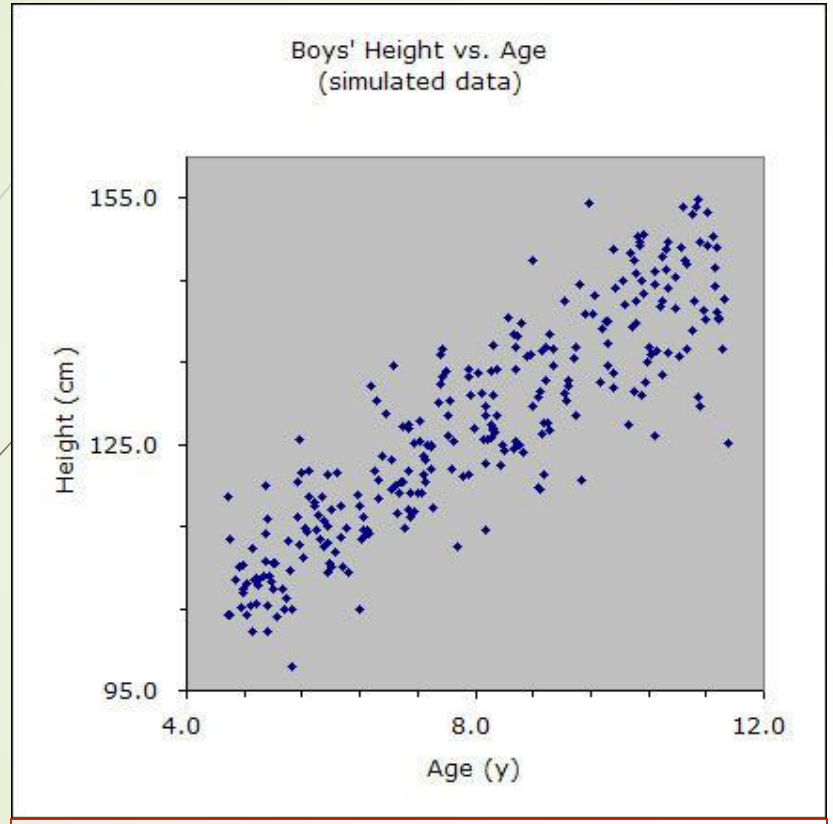
1: el experimentador fija los valores de las x_i y obtiene "al azar" los correspondientes y_i

2: el experimentador obtiene "al azar" parejas de valores (x_i , y_i)

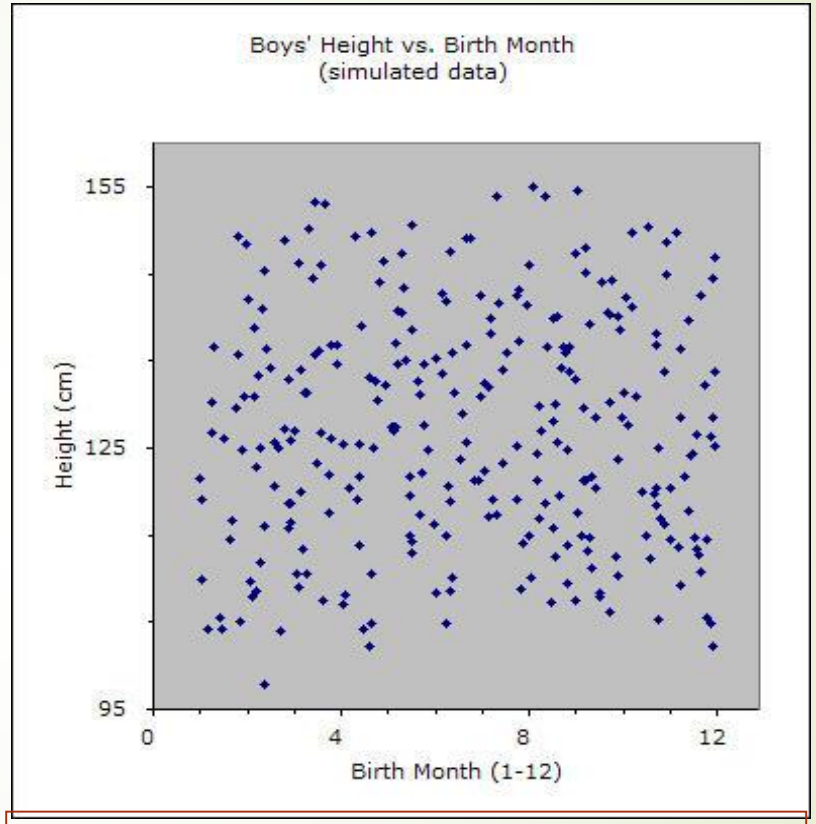
En ambos casos

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

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

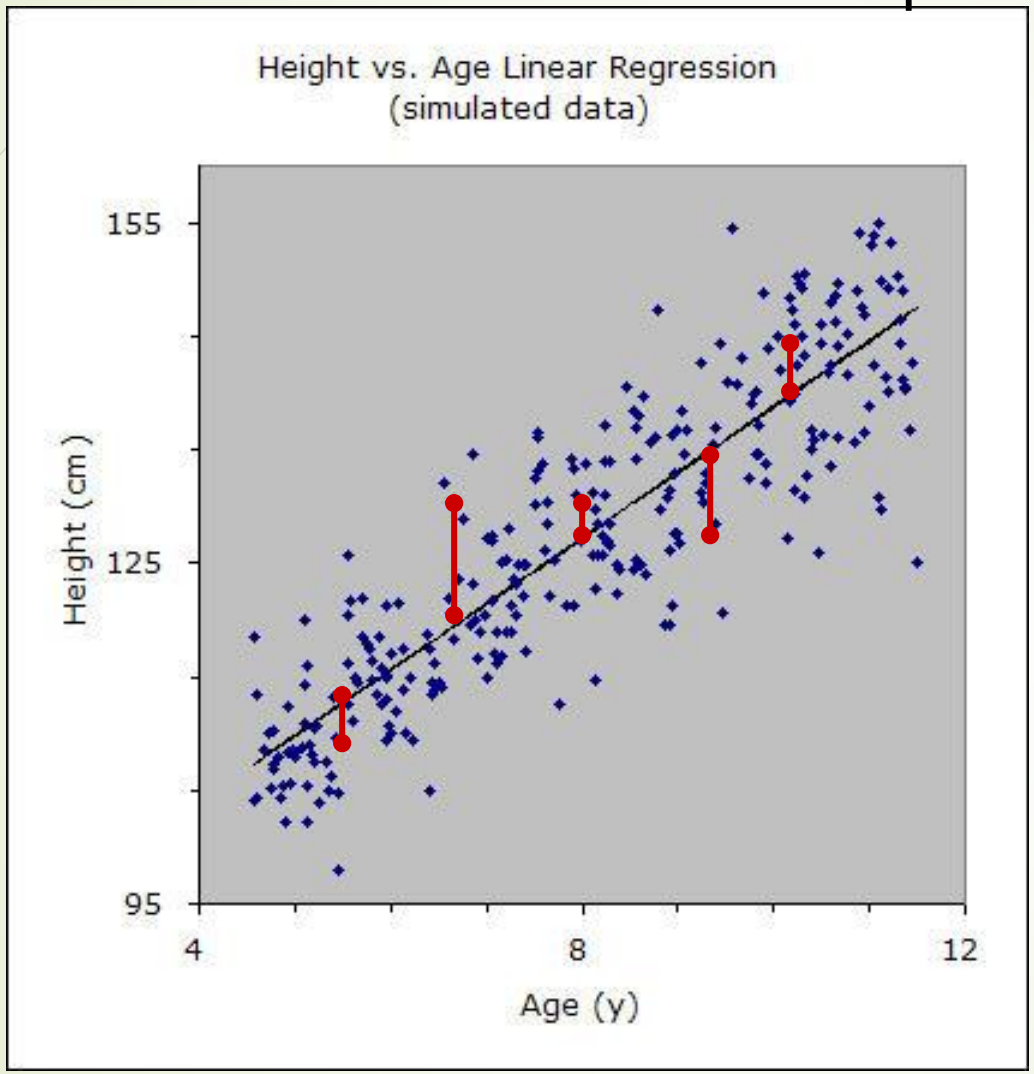


La nube de puntos permite suponer que la altura crece linealmente con la edad



¿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 (y) en función de la edad (x)?

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

Estimación de los parámetros de la regresión

- Parámetros de la recta: $\hat{\beta}_1 = \frac{cov_{xy}}{v_x}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Recta de regresión estimada $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Residuos (valores observados de la $\mathbf{U_x}$ en la muestra, n residuos)

Para cada (x_i, y_i) calculamos $u_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

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

$nv_y(1- r^2)$

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

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

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

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

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

Recordad: $nv_x = (n-1)\ s_x^2$

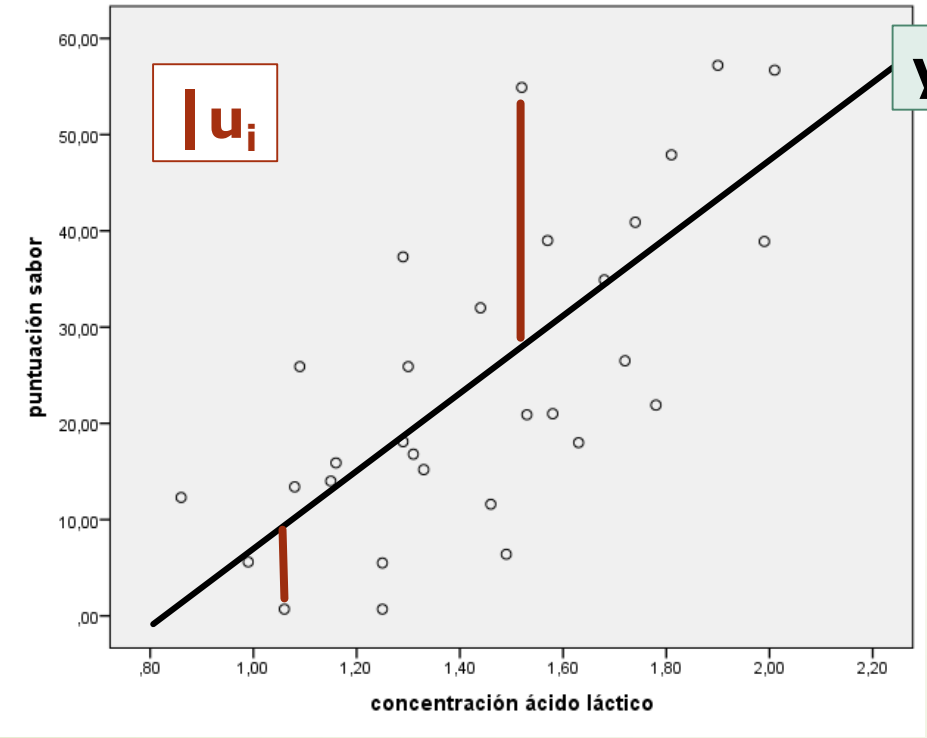
Ejemplo

X (láctico)	Y(sabor)
1,06	0,7
1,25	0,7
1,25	5,5
0,99	5,6
1,49	6,4
1,46	11,6
0,86	12,3
1,08	13,4
1,15	14,0
1,33	15,2
1,16	15,9
1,31	16,8
1,63	18,0
1,29	18,1
1,53	20,9
1,58	21.0
1,78	21,9
1,09	25,9
1,30	25,9
1,72	26,5
1,44	32,0
1,68	34,9
1,29	37,3
1,99	38,9
1,57	39,0
1,74	40,9
1,81	47,9
1,52	54,9
2,01	56,7
1,90	57,2

Queremos explicar el sabor de una porción de queso Cheddar en relación con su concentración de ácido láctico.

X= concentración de ácido láctico Y = puntuación del sabor

- **Modelo:** $Y_x = \beta_0 + \beta_1 x + U_x$ = puntuación del sabor en una porción de queso, elegida al azar, con una cantidad **x** de ácido láctico
- **Datos:** se analizan n = 30 porciones de queso (30 puntos en el plano)



y_x=-29,9+37,7x (recta ajustada)

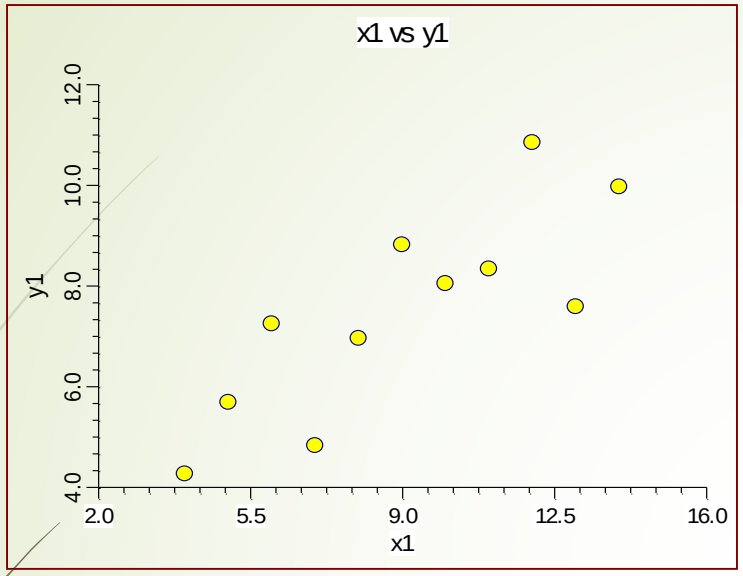
Resumen	Láctico	Sabor
Media	1,44	24,5
Varianza	0,09	255,4
Covarianza	3,358	
Estimación de β_1	**37,7**	
Estimación de β_0	**-29,9**	
Estimación de σ^2	**S_R^2 =137,95**	

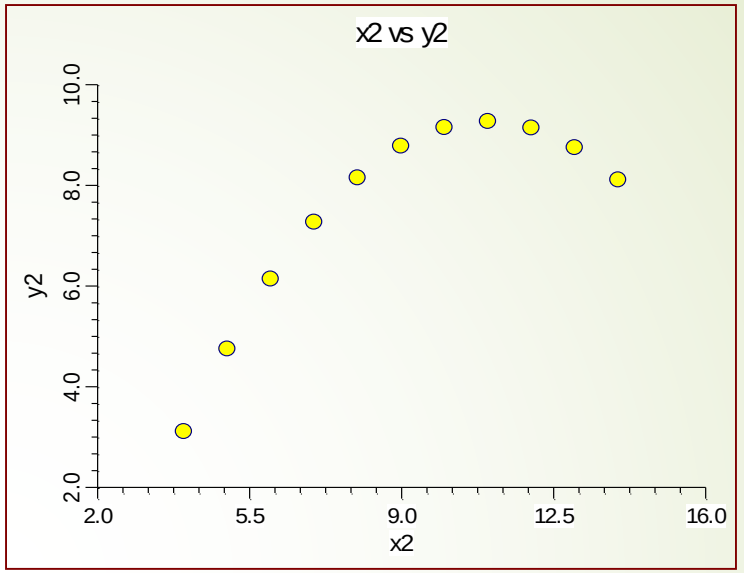
La importancia de los gráficos de puntos
(4 conjuntos de 11 datos emparejados)

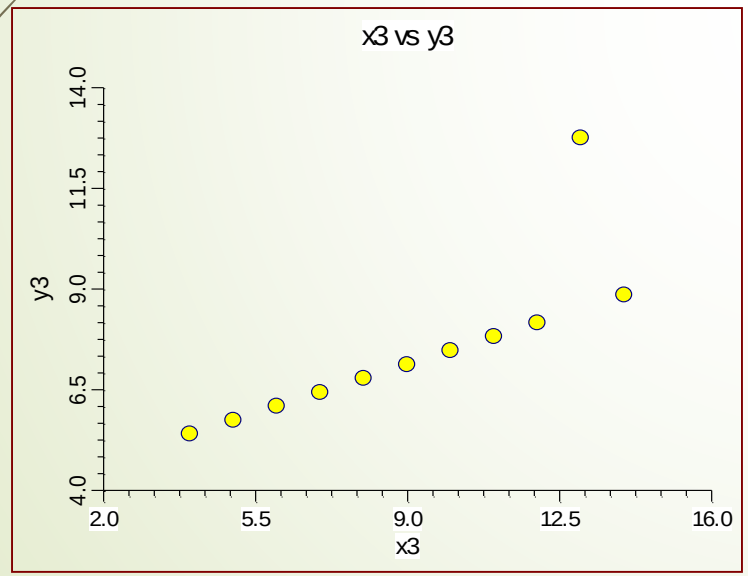
Mismos resultados resumidos

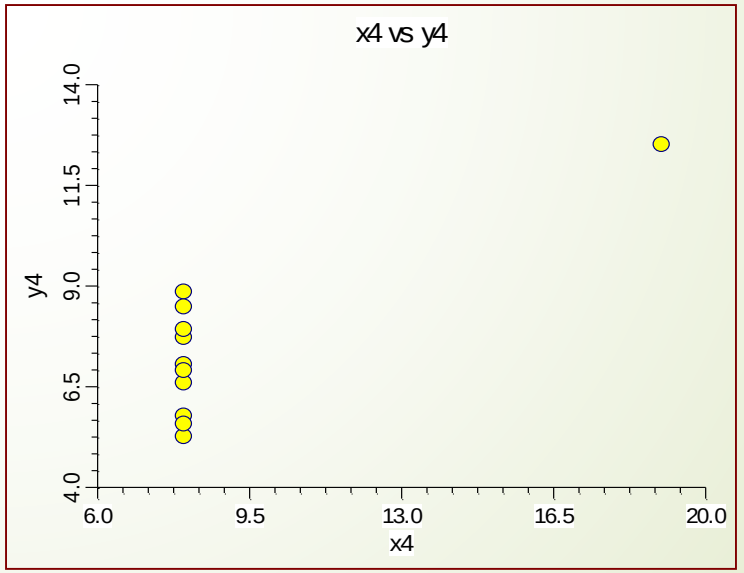
	datos 1		datos 2		datos 3		datos 4	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	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
Medias	**9**	**7,5**	**9**	**7,5**	**9**	**7,5**	**9**	**7,5**
Varianzas	**11**	**4,1**	**11**	**4,1**	**11**	**4,1**	**11**	**4,1**
Coef. correlación	**0,82**		**0,82**		**0,82**		**0,82**	
Recta de regresión	**y = 3 + 0,5x**		**y = 3 + 0,5x**		**y = 3 + 0,5x**		**y = 3 + 0,5x**	

Pero los gráficos son:









Análisis de los residuos

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

Los residuos pueden dibujarse de distintas formas:

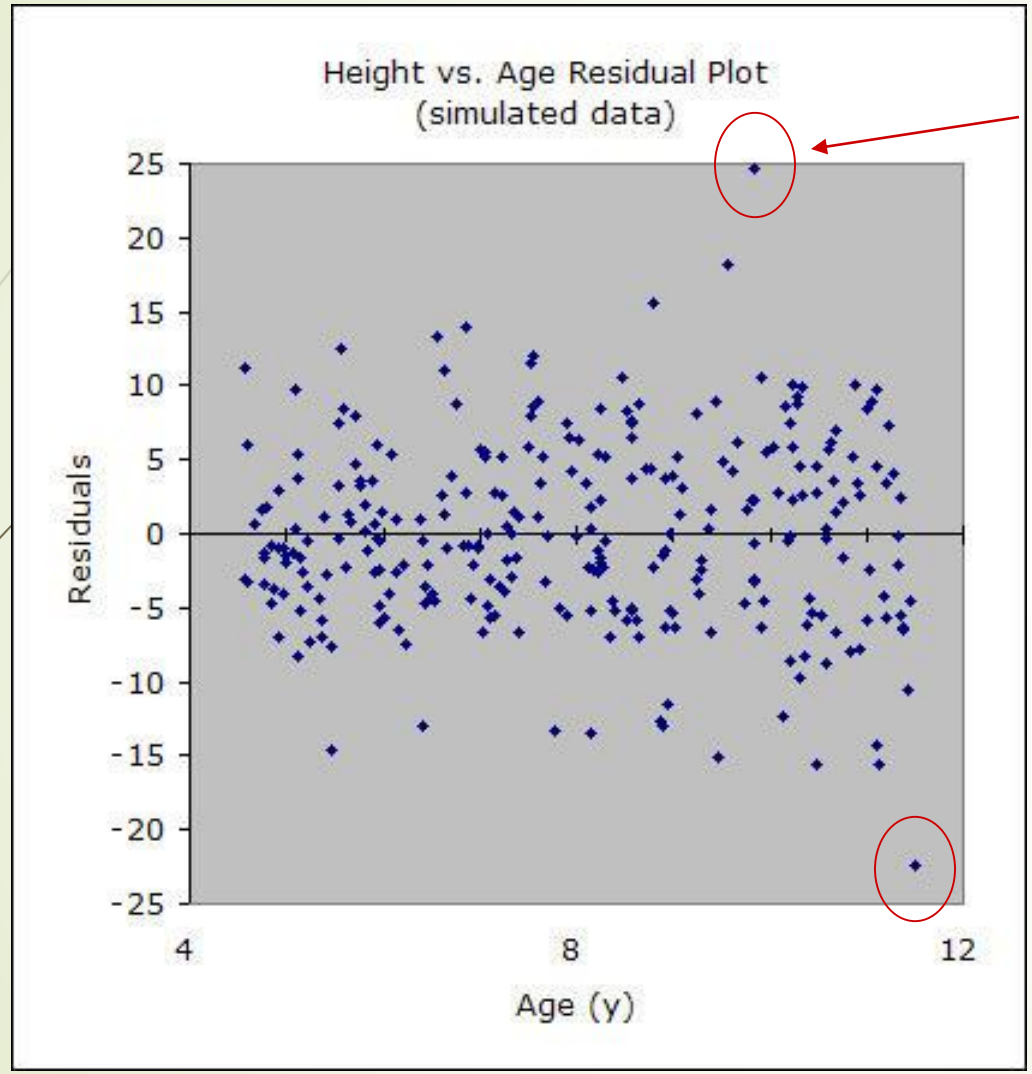
1. poniendo en el eje de abcisas los valores de las x_i y en ordenadas los correspondientes u_i
2. poniendo en el eje de abcisas los valores pronosticados de las y_i y en ordenadas los correspondientes u_i

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 sin forma alrededor del cero. Residuos tipificados muy alejados del cero (lejos de (-2,2) donde deben situarse el 95% de los puntos) pueden indicar datos anómalos.

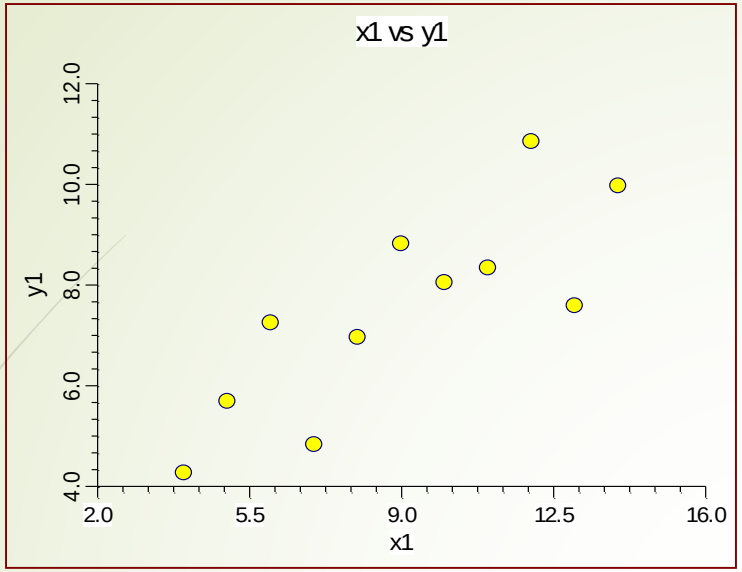
Gráfico de los residuos u_i

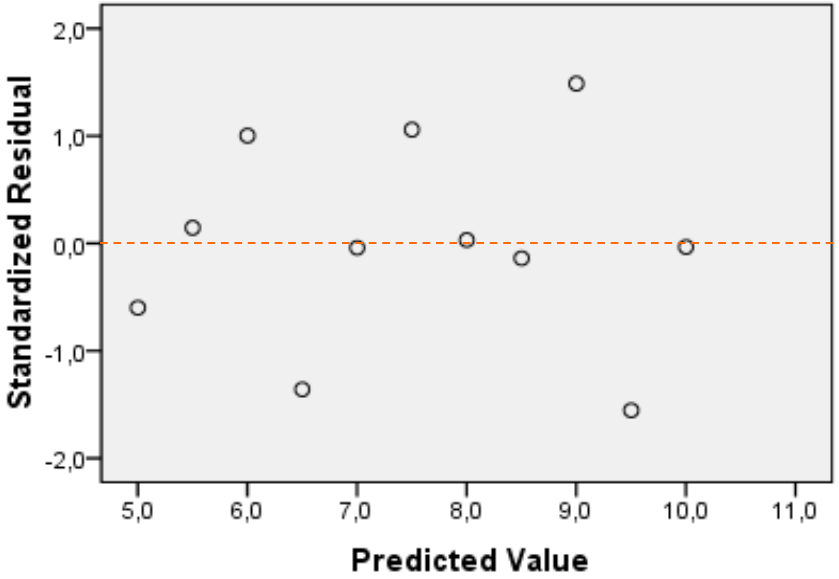


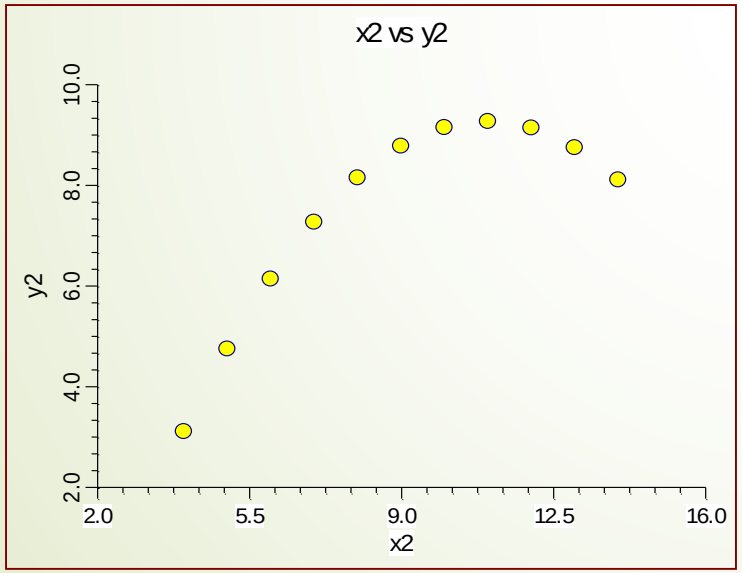
¿es este un valor anómalo?

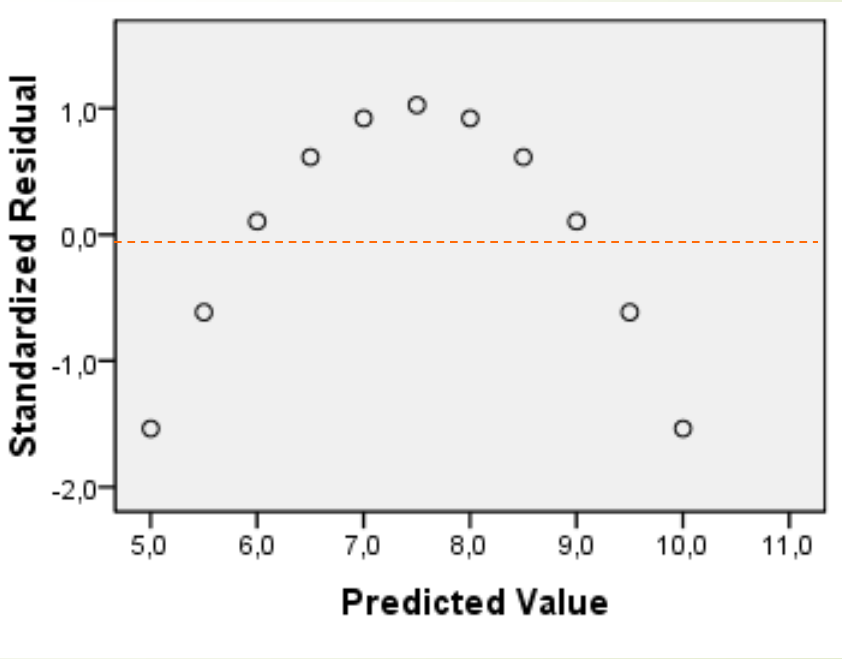
En abcisas los valores de x_i (edades en años)

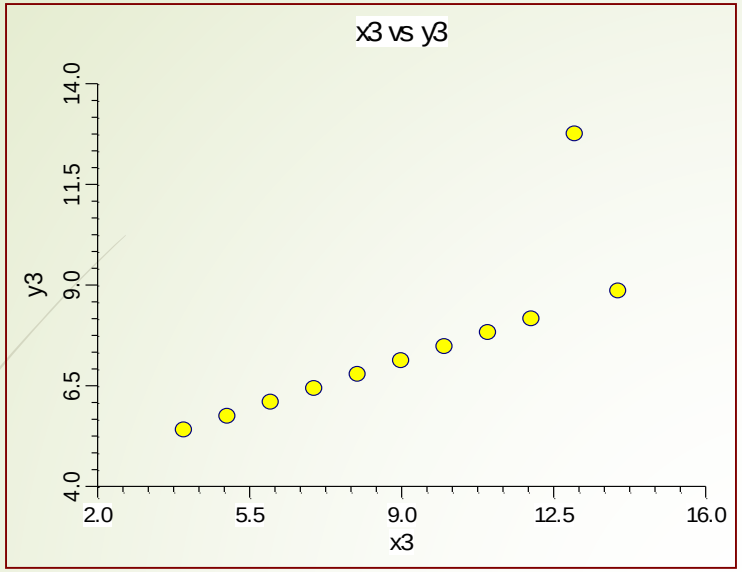
En ordenadas los residuos u_i sin tipificar

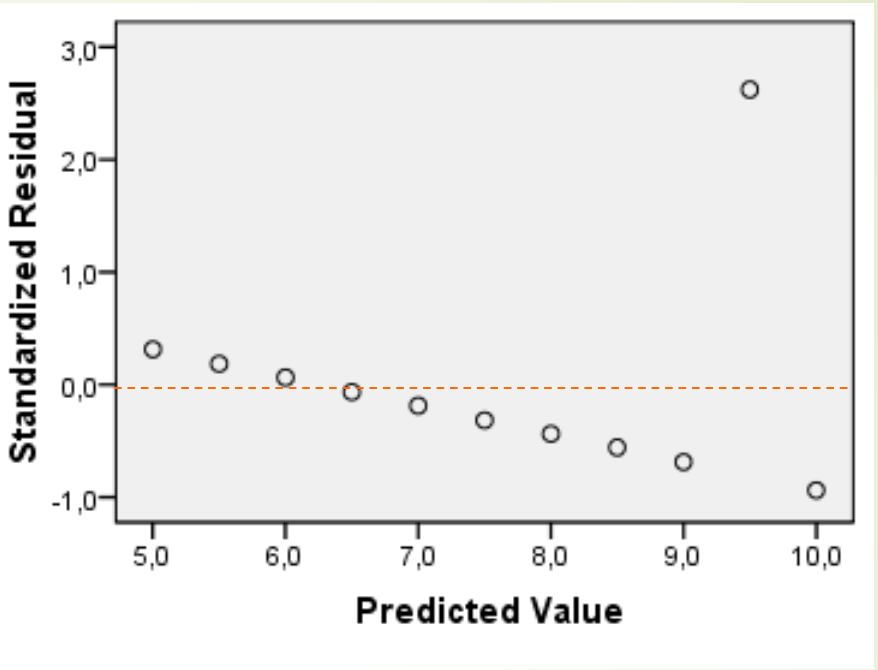

x1 vs y1
y1
x1
12.0
10.0
8.0
6.0
4.0
2.0
5.5
9.0
12.5
16.0

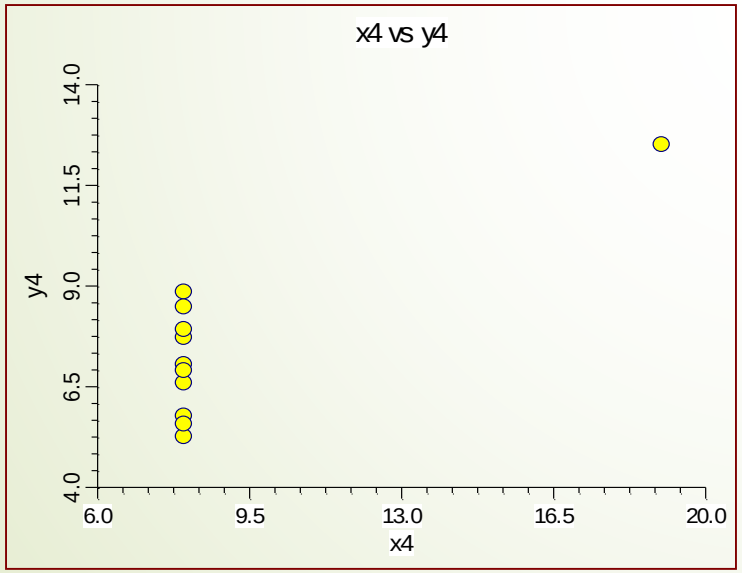

Standardized Residual
Predicted Value
2,0
1,0
0,0
-1,0
-2,0
5,0
6,0
7,0
8,0
9,0
10,0
11,0

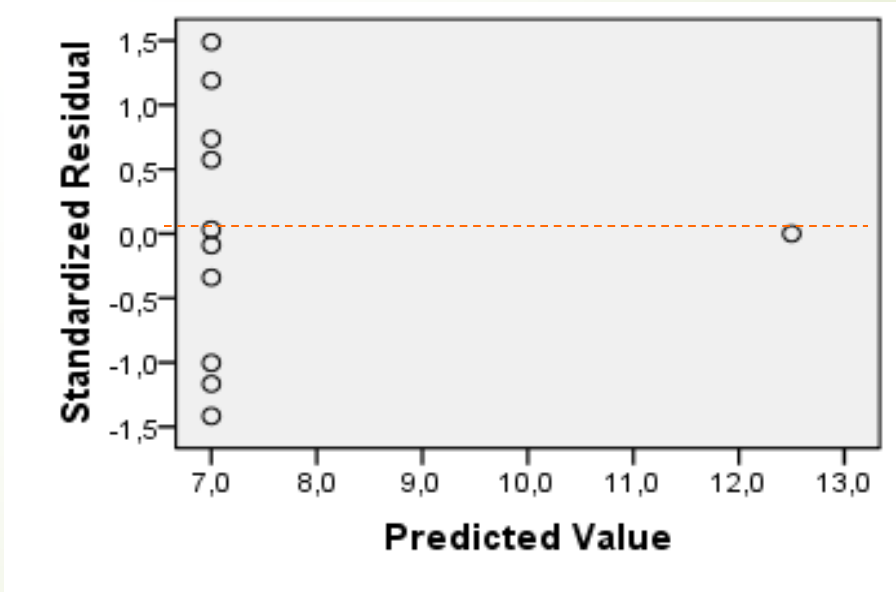

x2 vs y2
y2
x2
10.0
8.0
6.0
4.0
2.0
5.5
9.0
12.5
16.0


Standardized Residual
Predicted Value
1,0
0,0
-1,0
-2,0
5,0
6,0
7,0
8,0
9,0
10,0
11,0


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


Standardized Residual
3,0
2,0
1,0
0,0
-1,0
5,0
6,0
7,0
8,0
9,0
10,0
Predicted Value


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


Standardized Residual
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
7,0
8,0
9,0
10,0
11,0
12,0
13,0
Predicted Value

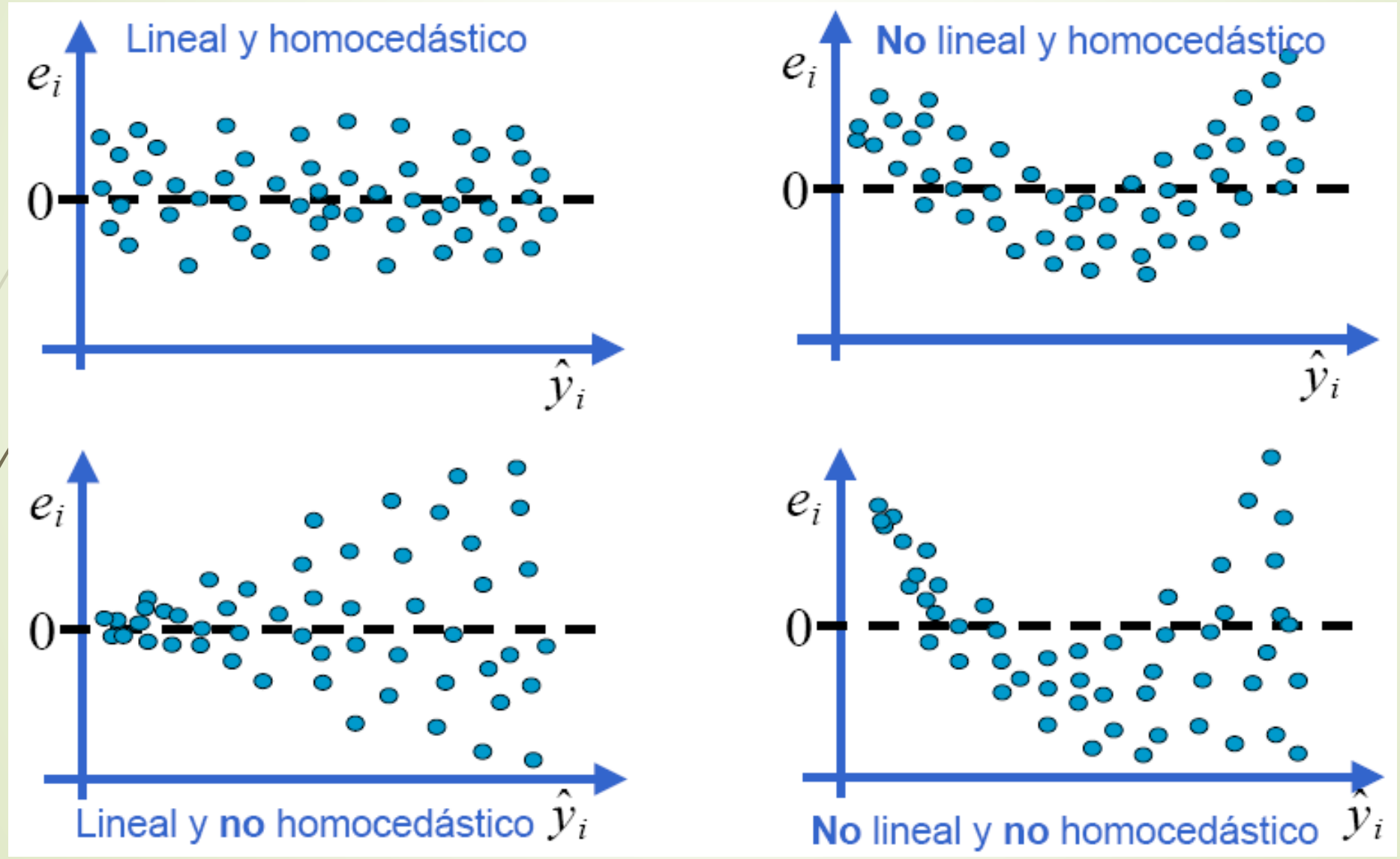
Requisitos del modelo

1. **Normalidad:** Los residuos se ajustan a una distribución Normal (histograma, gráfico P-P, contraste de K-S)
2. **Homocedasticidad:** la variabilidad de los residuos para los distintos valores de x es similar
3. **Linealidad:** los residuos 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

RESIDUOS vs VALORES PRONOSTICADOS

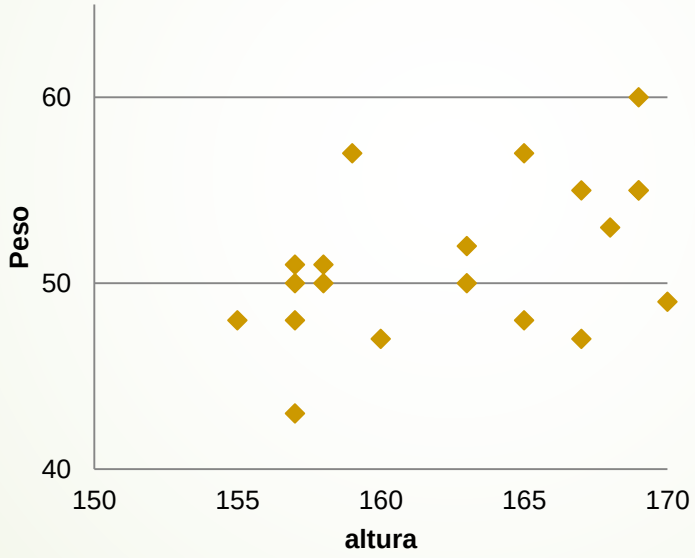
¿Se cumplen las hipótesis de linealidad e igualdad de varianzas?



Ejemplo 1: la siguiente tabla recoge los datos de altura (cm.) y peso (Kg.) de 20 mujeres estudiantes de la UAM

altura	peso
159	57
160	47
168	53
157	50
157	43
155	48
165	48
157	48
167	55
163	52
169	55
158	50
169	60
158	51
157	51
163	50
170	49
165	57
167	47
169	55

	estatura	*peso*
Media	162,65	51,30
Desviación típica	5,14	4,19
Varianza de la muestra	26,45	17,59



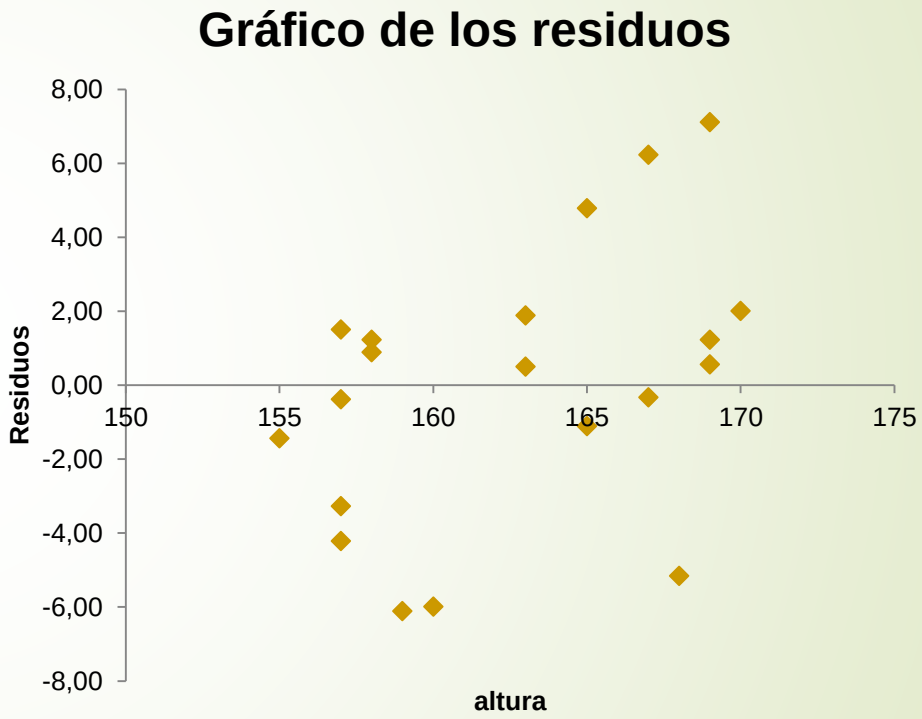
Coeficiente de correlación 0,476

Estimaciones

β_0	-11,84
β_1	0,388

Análisis de los residuos brutos

altura	peso	Pronóstico peso	Residuo
159	57	49,88	7,12
160	47	50,27	-3,27
168	53	53,38	-0,38
157	50	49,11	0,89
157	43	49,11	-6,11
155	48	48,33	-0,33
165	48	52,21	-4,21
157	48	49,11	-1,11
167	55	52,99	2,01
163	52	51,44	0,56
169	55	53,77	1,23
158	50	49,49	0,51
169	60	53,77	6,23
158	51	49,49	1,51
157	51	49,11	1,89
163	50	51,44	-1,44
170	49	54,15	-5,15
165	57	52,21	4,79
167	47	52,99	-5,99
169	55	53,77	1,23



TRANSFORMACIONES DE LOS DATOS

Cuando detectamos problemas de

no linealidad
o
heterocedasticidad

Y queremos aplicar las técnicas de regresión lineal

Algunas funciones linealizables

Exponencial $y = ke^{\beta x}$ —Log→ $logy = logk + \beta x$

Potencial $y = kx^{\beta}$ —Doble Log→ $logy = logk + \beta logx$

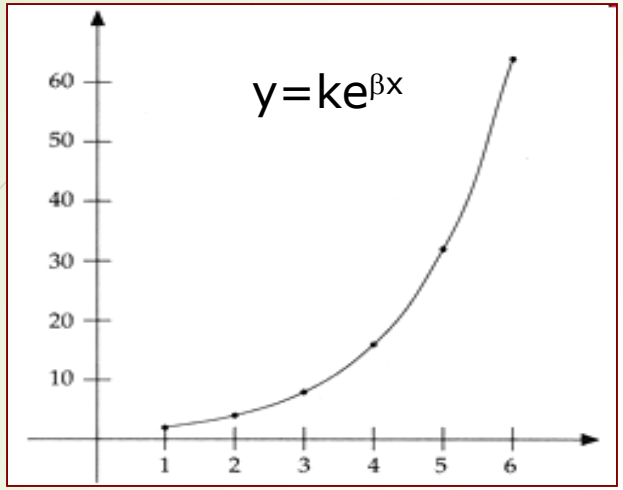
Logarítmico $y = \beta_0 + \beta_1 logx$

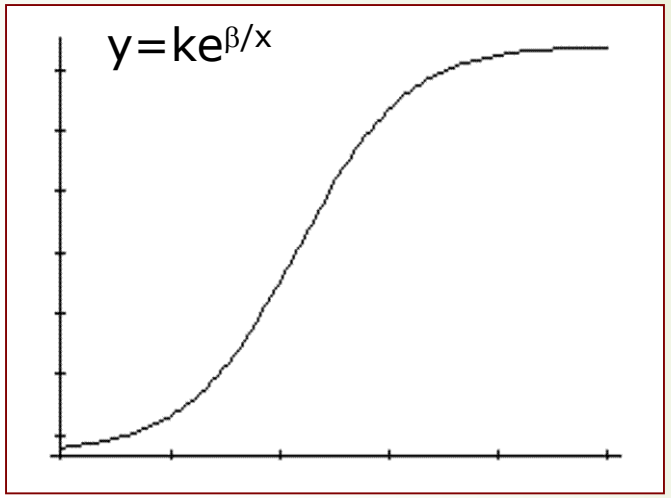
Inversa $y = k + \beta\frac{1}{x}$

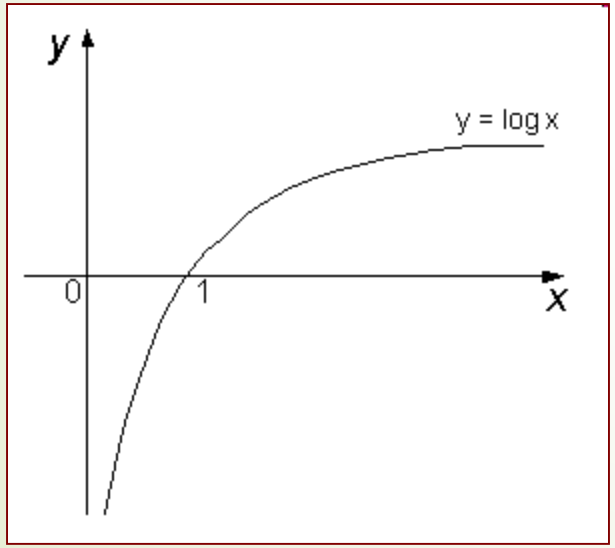
Mixtos

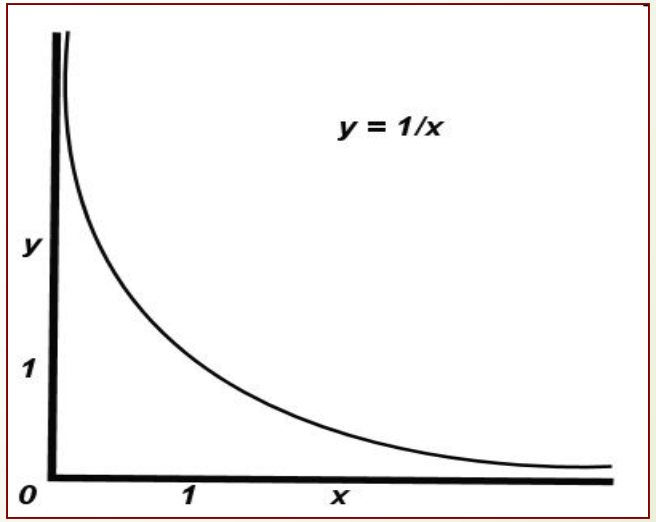
$y = ke^{\frac{\beta}{x}}$ —Log + 1/x→ $logy = logk + \beta\frac{1}{x}$

Algunas gráficas

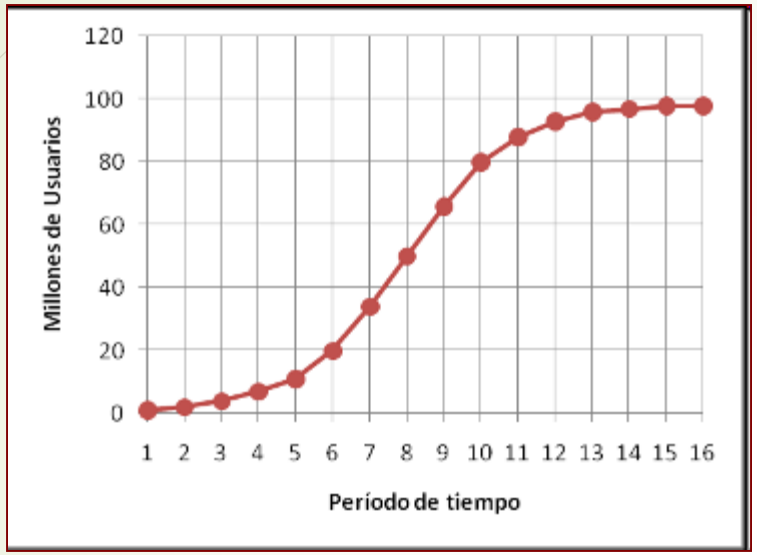








La curva logística



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

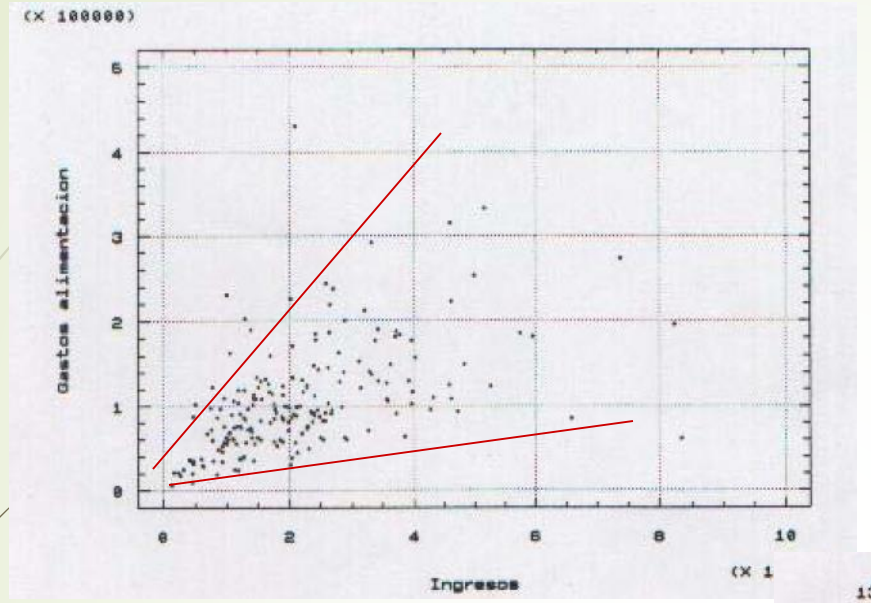
Nota: C es el valor máximo posible de la variable Y

Cambio de variable:

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

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

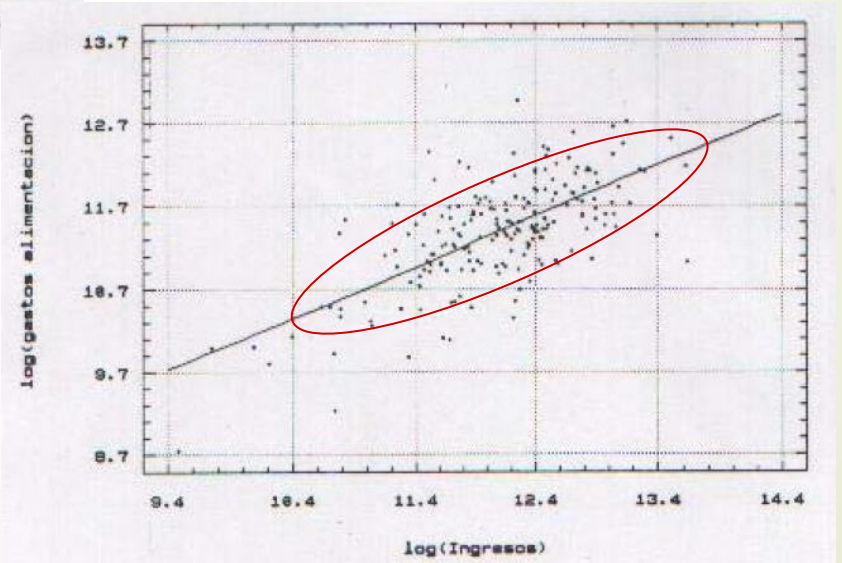
Problemas de Heterocedasticidad



Transformamos

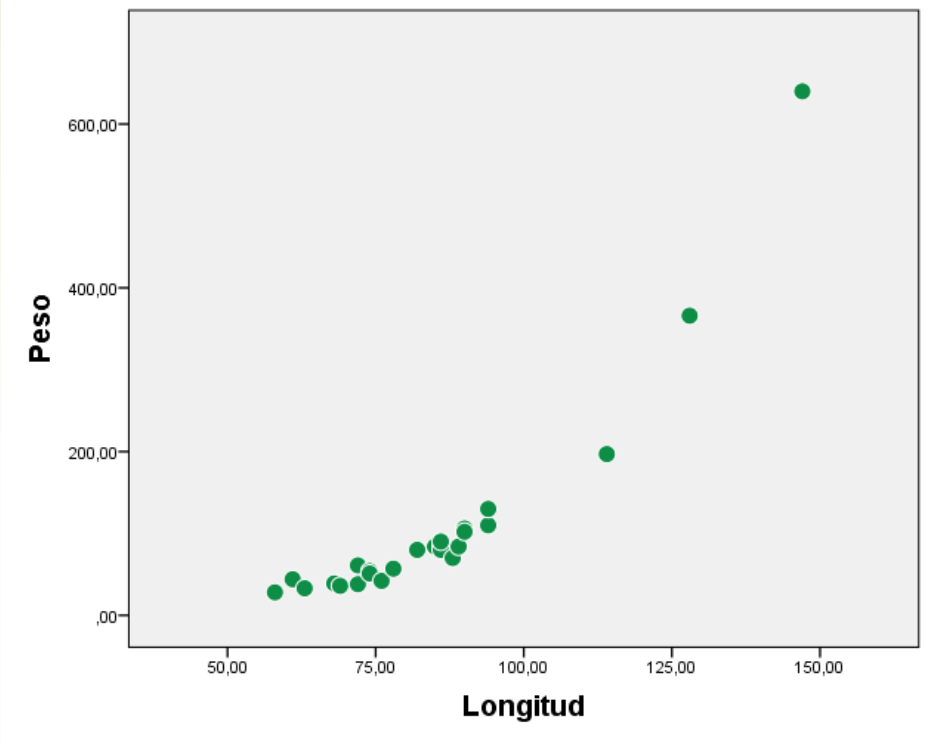
Transformación doble log

Transformamos



Ejemplo 2. 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:

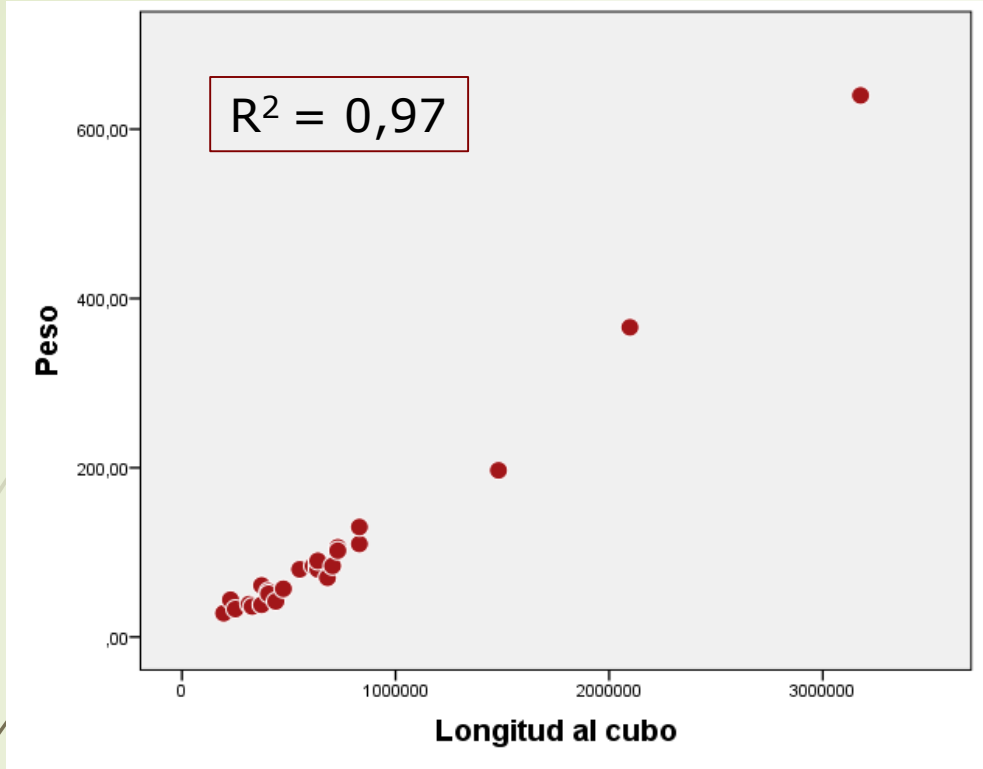


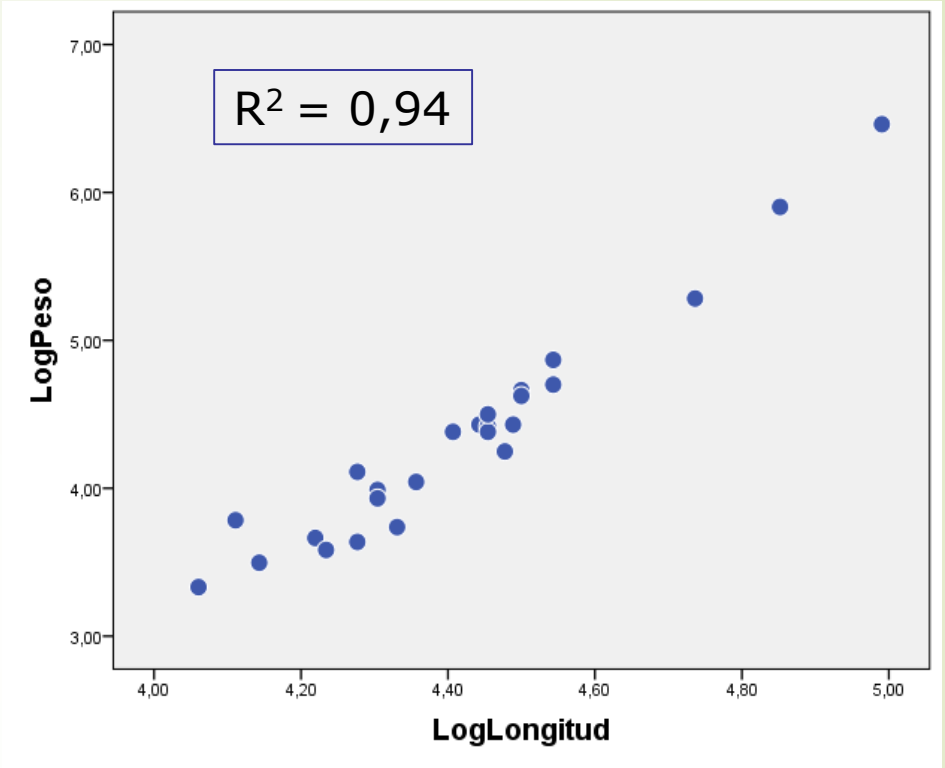
¿Qué función representa mejor el peso (Y) en función de la longitud (X)?

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

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

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





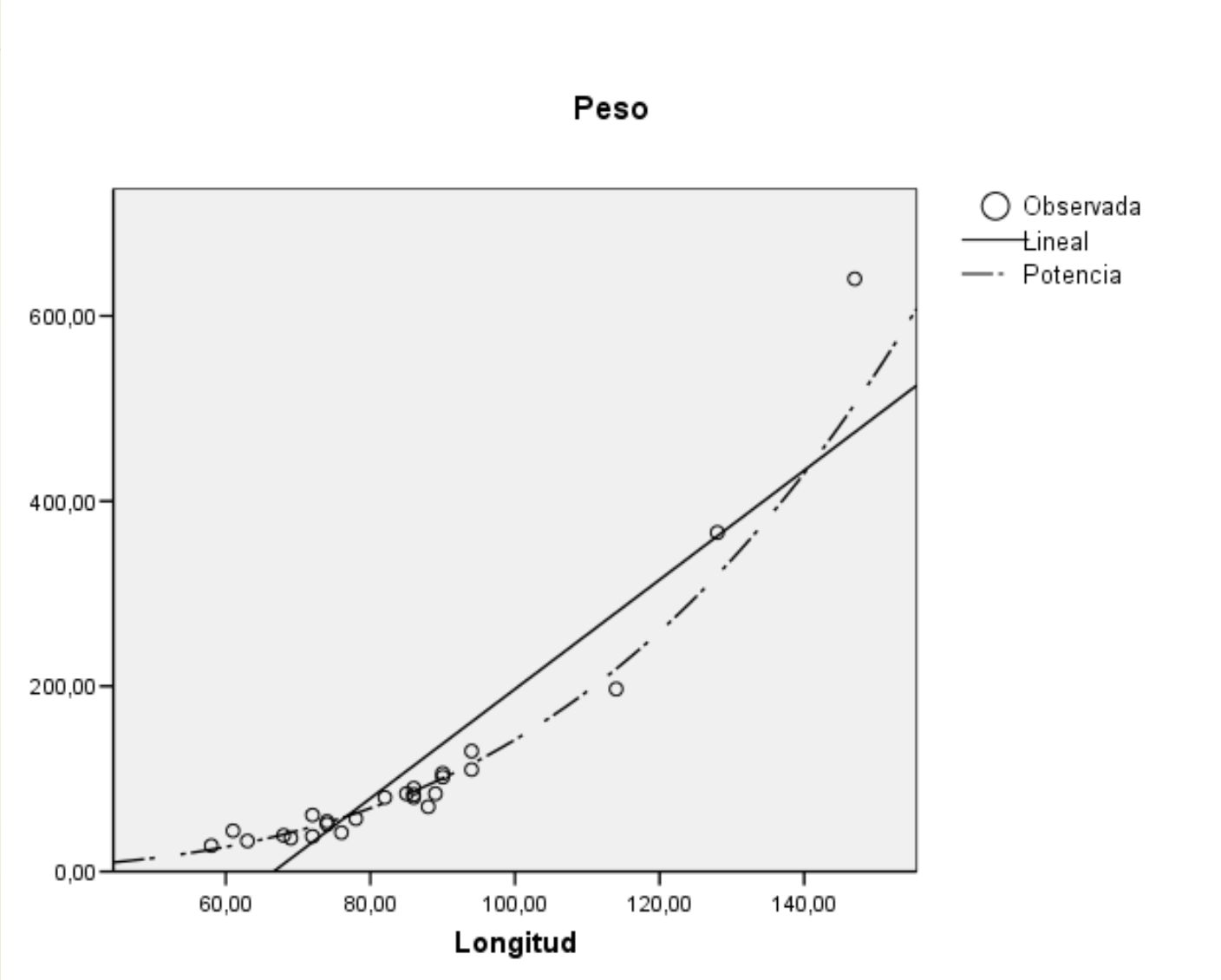
¿Qué modelo representa cada una de estas gráficas?

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

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

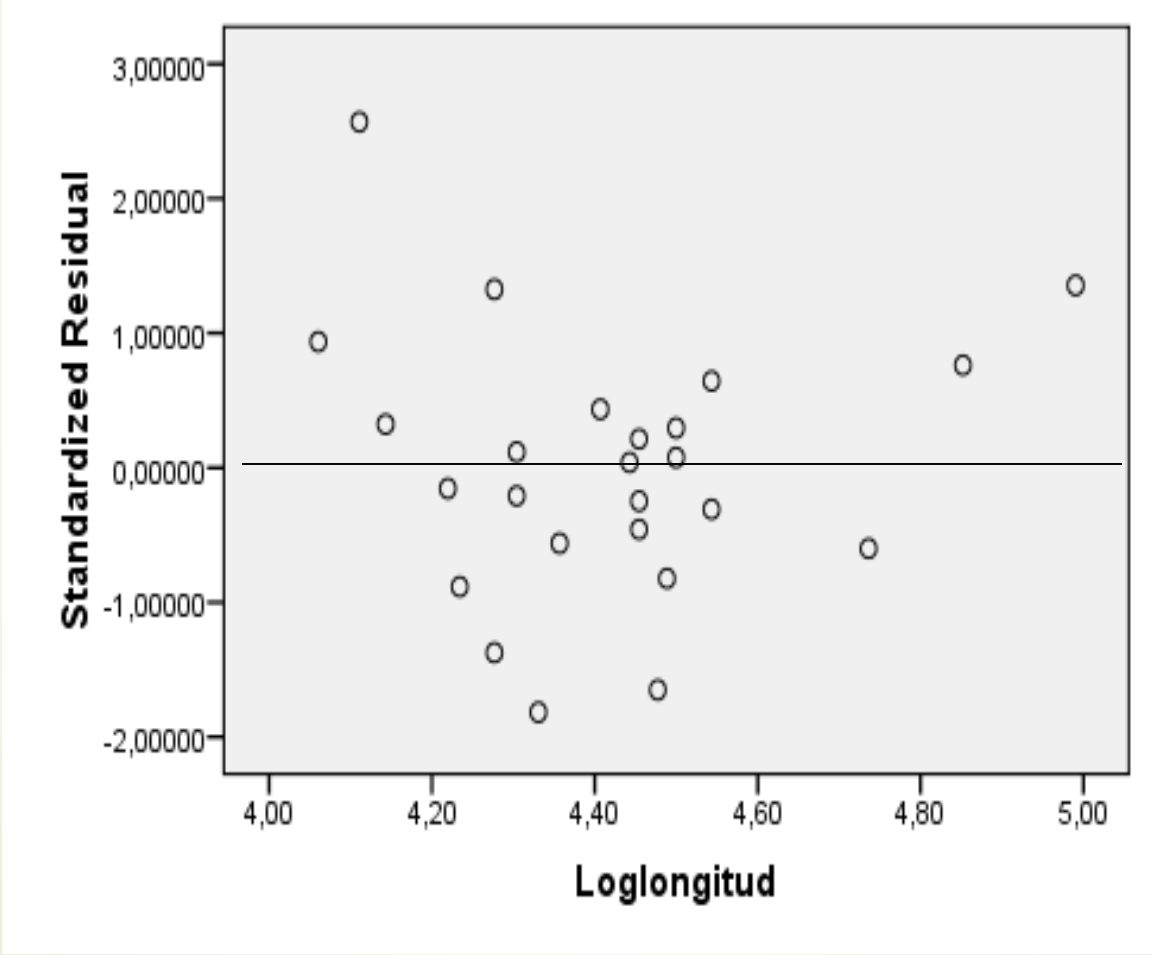
Modelo $Y = k X^{\beta 1}$

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

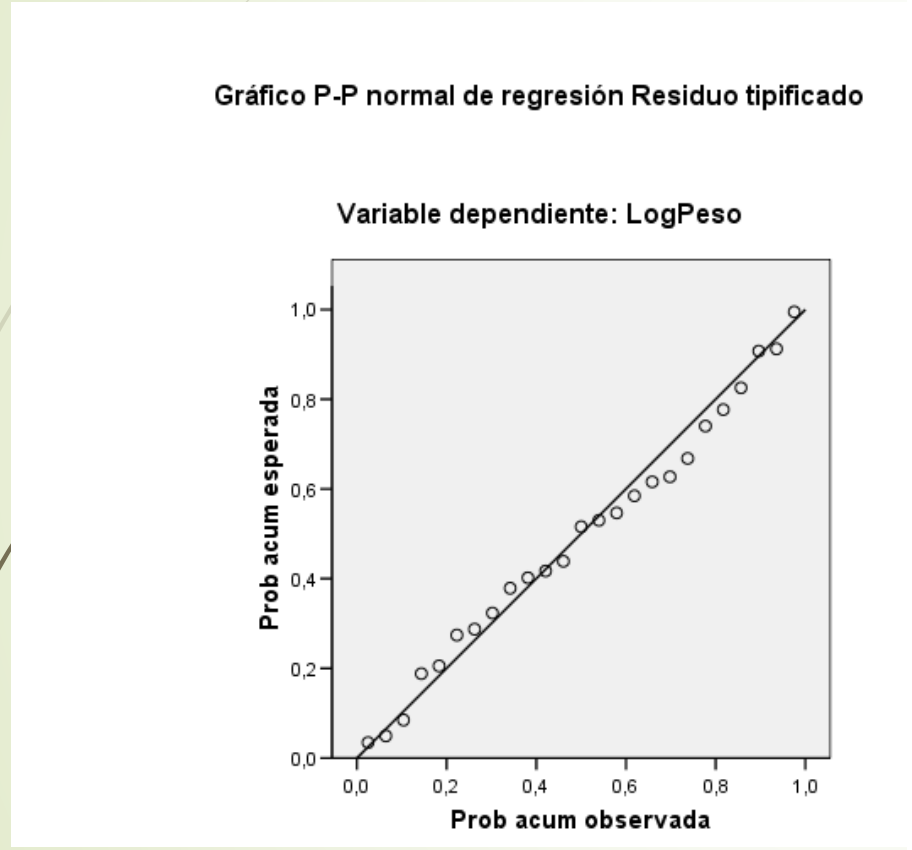


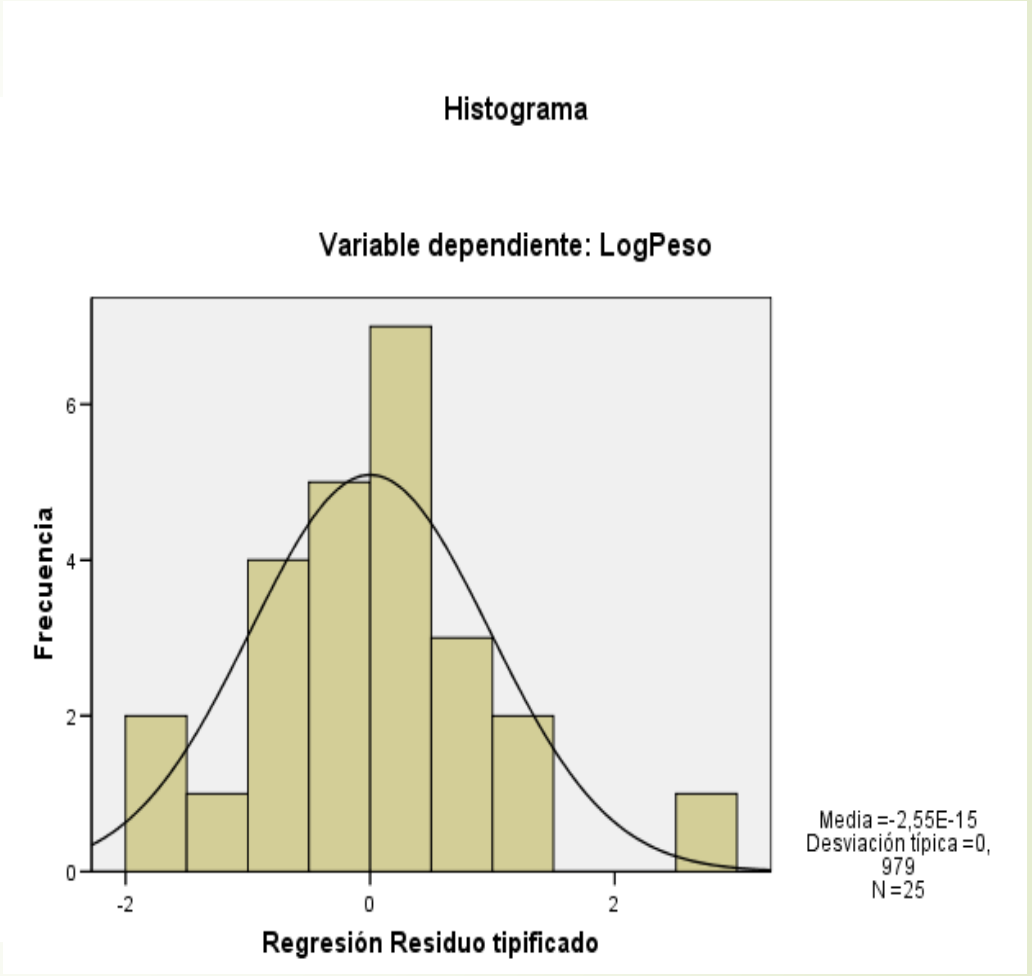
Modelo $Y = k X^{\beta 1}$: análisis de los residuos

Residuos tipificados sobre Log(longitud)



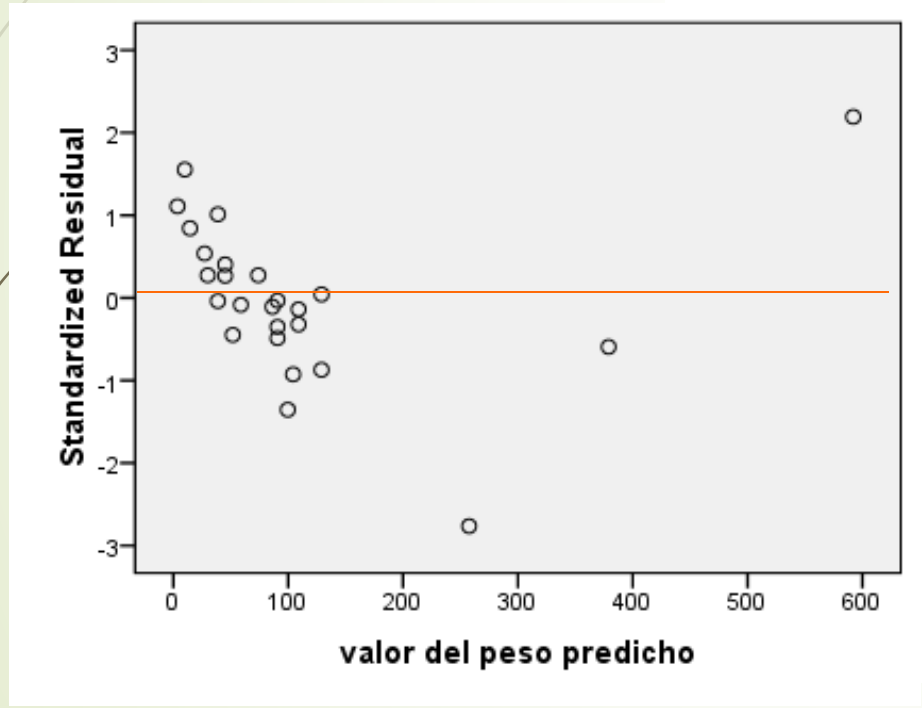
Modelo $Y = k X^{\beta 1}$: análisis de los residuos (Normalidad)

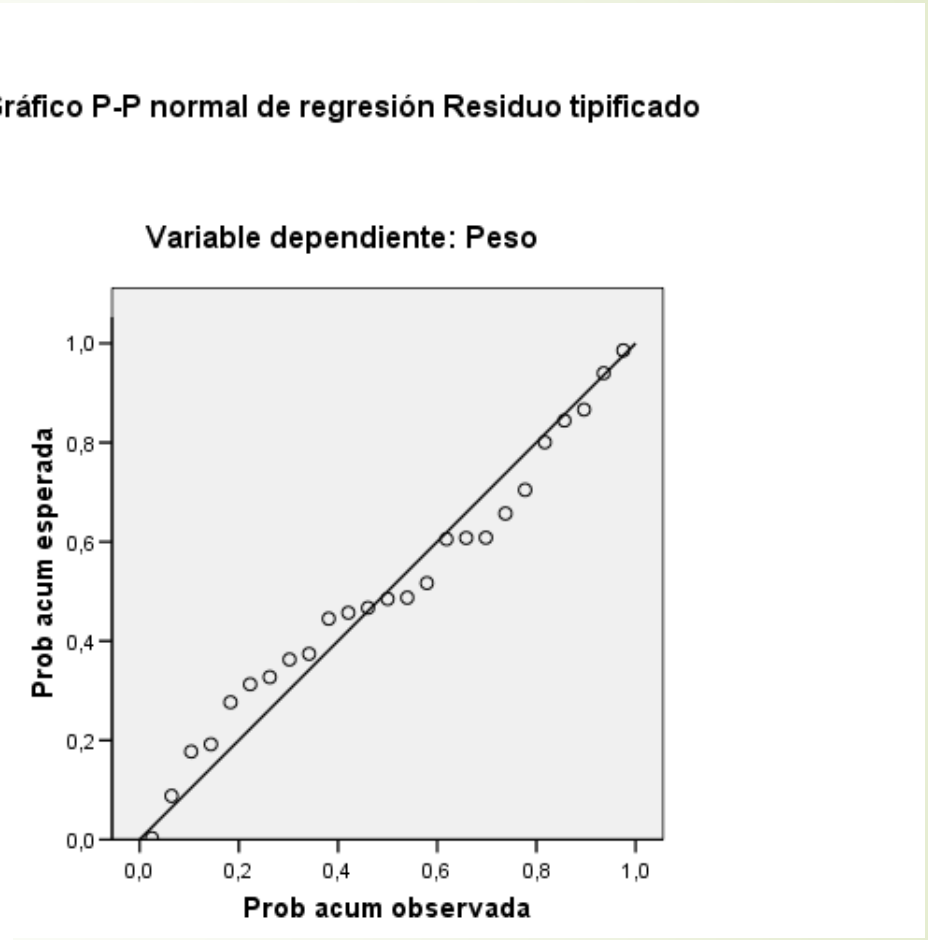




Modelo $Y = \beta_0 + \beta_1 X^3$: análisis de los residuos

Residuos tipificados





Contraste de regresión

$$Y_x = \beta_0 + \beta_1 x + U_x$$

$H_0 : \beta_1 = 0$ (la X no influye linealmente sobre la Y; el modelo NO es válido)

$H_1 : \beta_1 \neq 0$ (la X influye linealmente sobre la Y; el modelo es válido)

Tres maneras equivalentes de resolverlo:

- Mediante un **intervalo de confianza para β_1** (rechazamos H_0 al nivel α si el 0 no pertenece al intervalo

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

- Mediante la tabla de coeficientes (**contraste de la t** y p-valor)
- Mediante la tabla ANOVA (**contraste de la F** y p-valor)

En el ejemplo del queso Cheddar

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

Error típico del estimador de β_1

$IC_{0,95}(\beta_1)$ = **37,7** +- $t_{28;\ 0,025}$* **7,19** = 37,7 +- 2,048* 7,19 = **[23 ; 52]**

Tabla de coeficientes con Excel

Excel	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-29,859	10,58	-2,82	0,00869	-51,536	-8,182
Lactic	**37,799**	**7,186**	5,2488	1,4 E-05	**22,999**	**52,44**

Ejemplo 2. 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

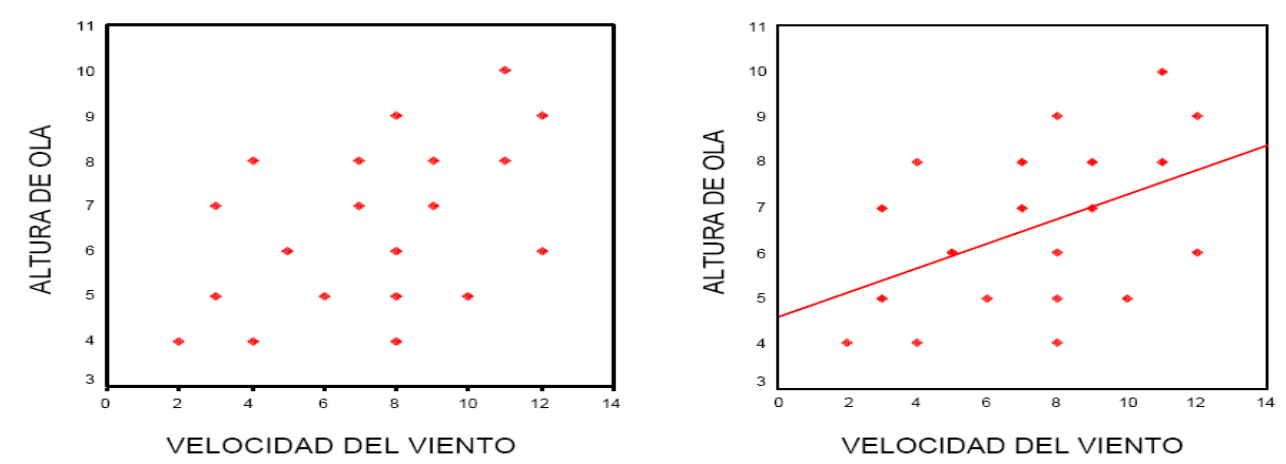
Curva de regresión estimada:

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

o equivalentemente:

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

Ejemplo 3. 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

CONTRASTES DE LA REGRESIÓN: ANOVA

Descomposición de la variabilidad en regresión

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

SCT = SCE + SCR

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

Tabla ANOVA

FV	SC	gl	Varianzas	F
Explicada por regresor	SCE	1	SCE	$F = \frac{\text{SCE}}{s_R^2}$
Residual	SCR	$n-2$	s_R^2	
Total	SCT	$n-1$		

- Rechazaremos **$H_0 : \beta_1 = 0$ (el modelo no es válido)** si

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

- Los ordenadores proporcionan la tabla y el **p-valor**
- Fórmulas útiles para el cálculo:

$$SCE = nv_y r^2 \quad SCR = nv_y(1- r^2) \quad SCT = nv_y$$

- Coeficiente de determinación $R^2 = SCE/SCT = r^2$

Ejemplo con ordenador (X= longitud, Y = anchura de una lapa)

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	13,020	1	13,020	76,423	,000[b]
	Residual	4,430	26	,170		
	Total	17,450	27			

a. Variable dependiente: Longitud

b. Variables predictoras: (Constante), Anchura

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	1,361	,468		2,907	,007	,399	2,323
	Anchura	1,996	,228	,864	8,742	,000	1,527	2,466

a. Variable dependiente: Longitud

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 explicativa

Ejemplo 3. 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 2. 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

Predicciones

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

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

  Estimación de la altura media de todos los posibles hijos de un padre que mide x_0

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

  Predicción de la altura de un hijo cuyo padre mide x_0
- En ambos casos la estimación puntual es la misma:

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

¿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 a una velocidad del viento fija x_0

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

El error que podemos cometer será mayor en el segundo caso
(mayor variabilidad en un valor que en la media)

Errores de predicción

- En el caso de la **Estimación de la media de Y dado X=x_0**

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

- En el caso de la **Predicción de un valor de Y dado X=x_0**

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

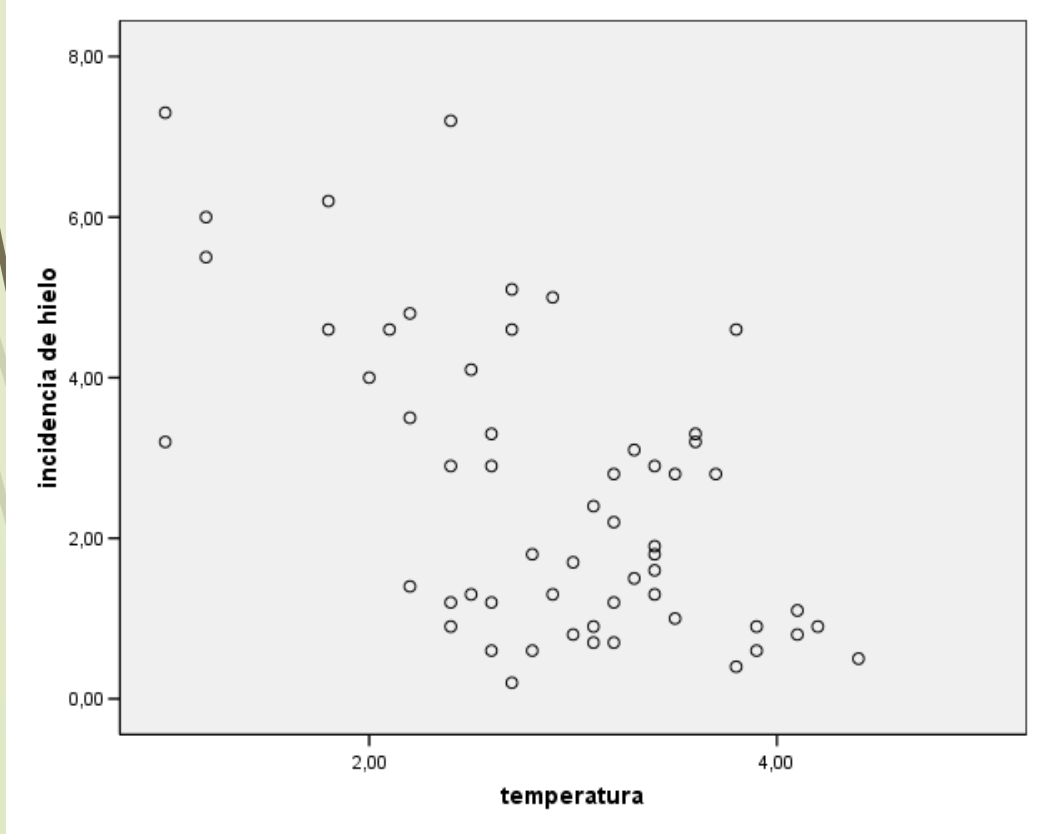
Error típico

¿Para qué valores de x_0 será menor el error típico?

Ejemplo

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

Datos de 57 años



	Mean annual temperature (°C)	Sea-ice incidence (months/year)
Medias	2,895	2,556
Varianzas	0,614	3,346
n	57	57
Covarianza	-0,852	
r	-0,595	

Ajuste del modelo lineal

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

	x_i	y_i	Pronóstico	**Residuo**
1	4,4	0,5	0,47	0,03
2	4,1	0,8	0,88	-0,08
3	4,2	0,9	0,74	0,16
4	4,1	1,1	0,88	0,22
56	2,4	7,2	3,24	3,96
57	1,0	7,3	5,19	2,11

Normalidad

Histograma

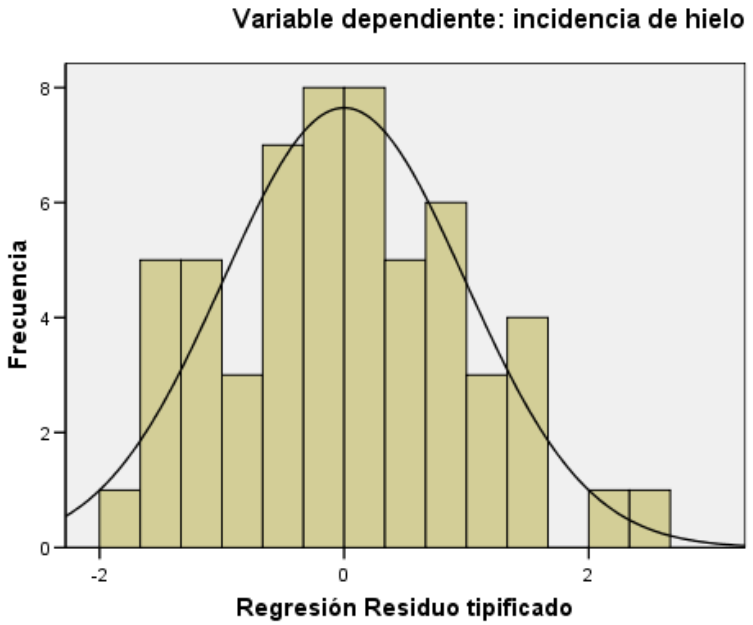
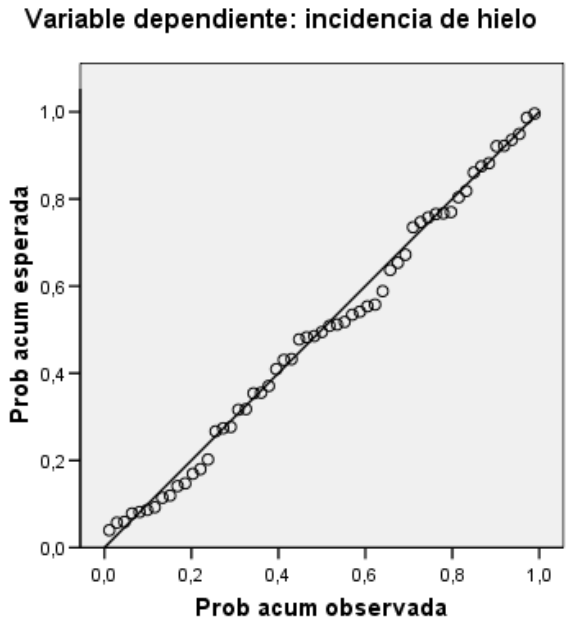
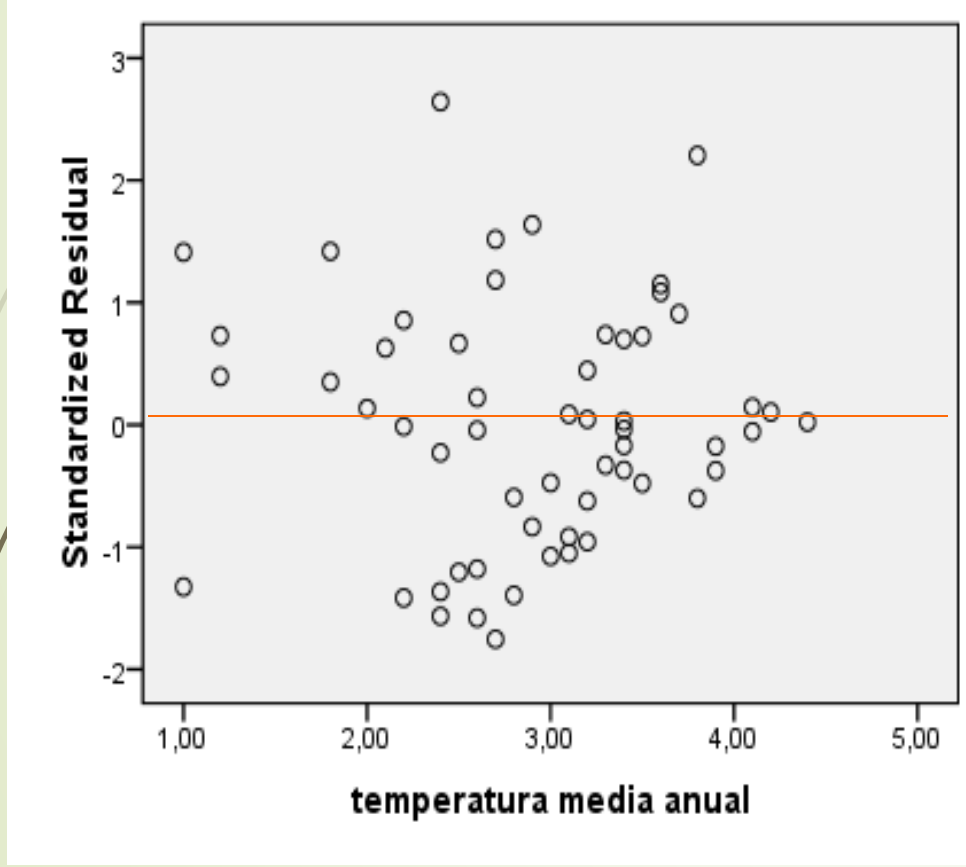
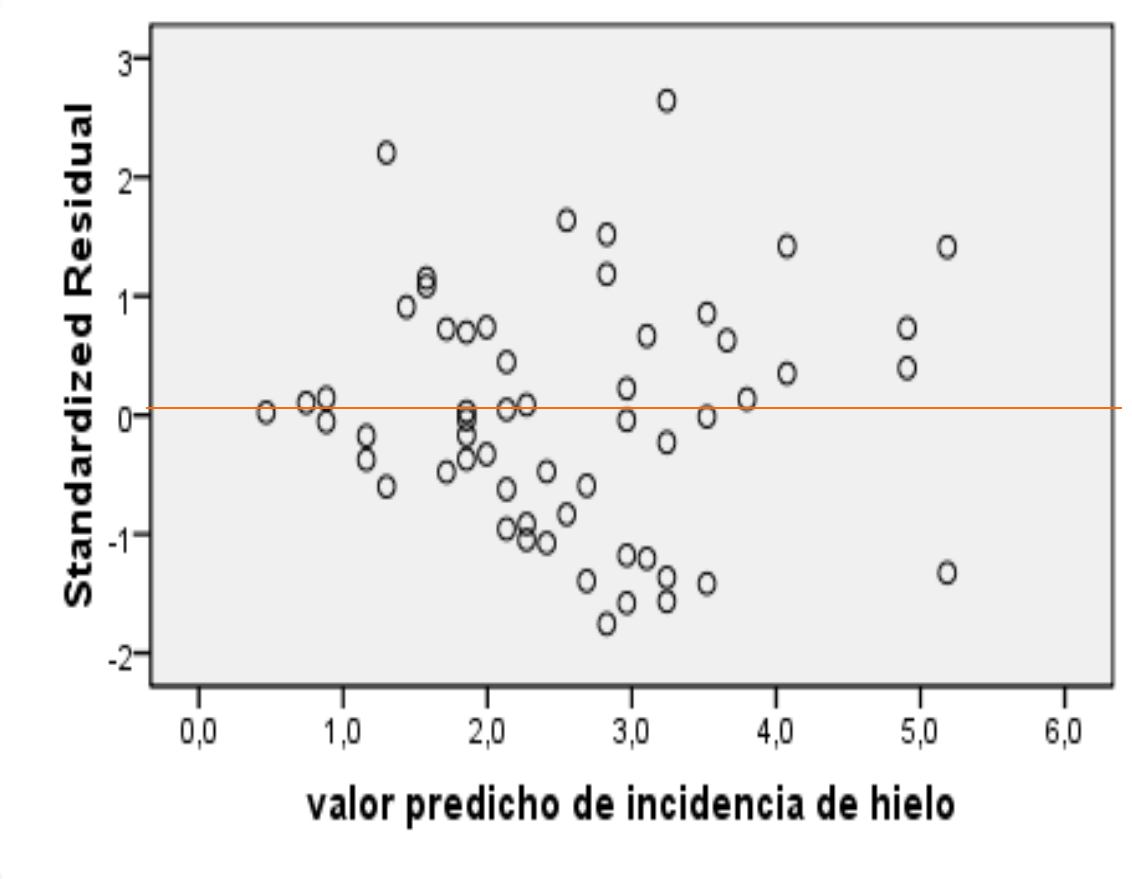


Gráfico P-P normal de regresión Residuo tipificado

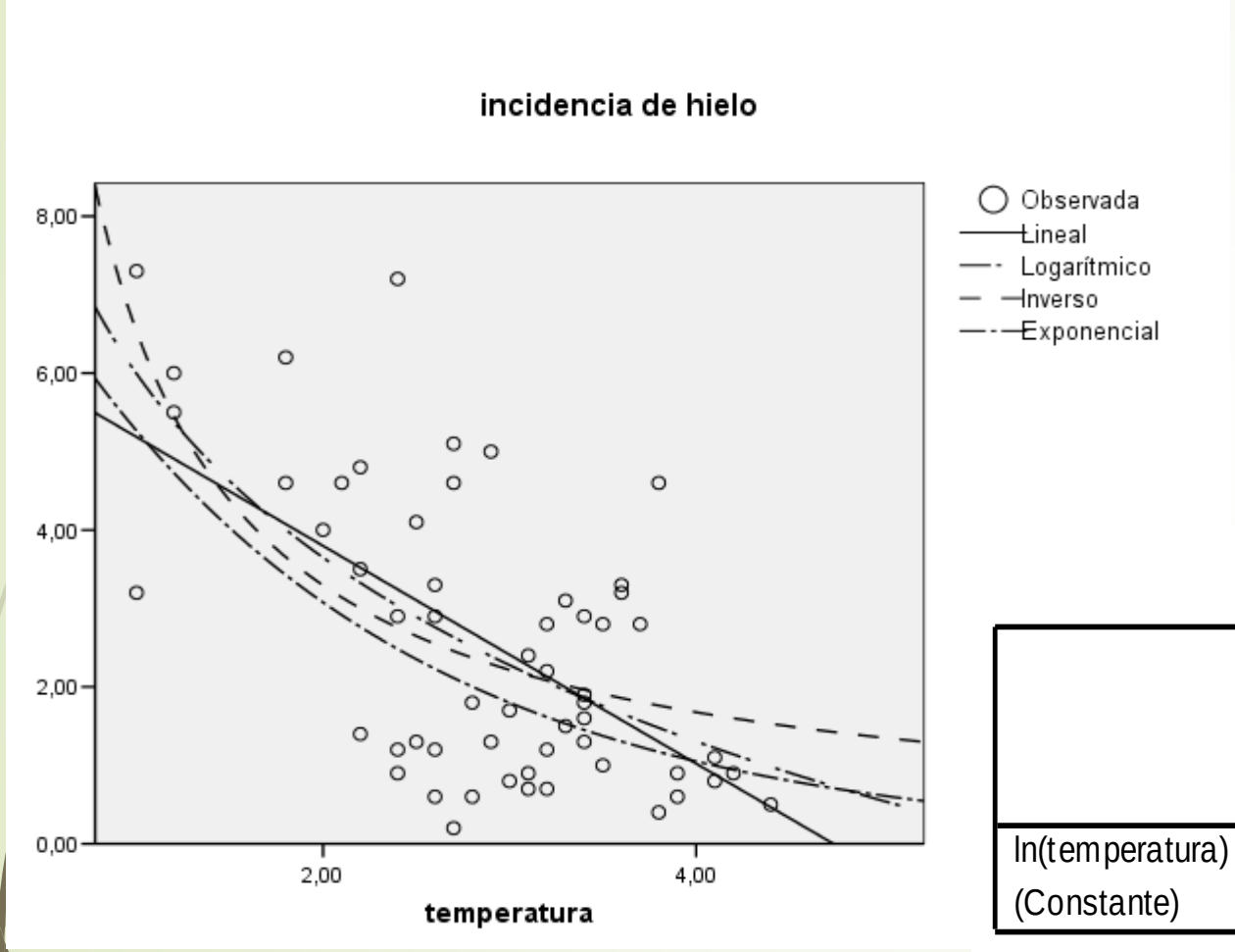


Gráficos de los residuos tipificados (linealidad, igualdad de varianzas, datos anómalos)





Otros modelos: transformaciones



Valores de r

Lineal:	r = -0,595
Logarítmico:	r = -0,609
Exponencial:	r = -0,514
Inverso:	r = -0,586

Coeficientes

	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 $y_x = 6{,}573 - 1{,}388\ x$

$$y_1 = 6{,}573 - 1{,}388 = 5{,}185 \text{ meses al año}$$

Intervalo de confianza 0,95 para la incidencia media de hielo:

$$5{,}185 \text{ -+ } t_{55,0.025}\ 1{,}497\ (0{,}515) = 5{,}185 \text{ -+ } 1{,}03 = (4{,}155,\ 6{,}215)$$

Respuesta con el modelo logarítmico: $y_x = 5{,}993 - 3{,}382 \log(x)$

$$y_1 = 5{,}993 - 3{,}382 \log(1) = 5{,}993 \text{ meses}$$

Con el modelo lineal $y_x = 6{,}573 - 1{,}388\ x$

¿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

y_x= incidencia de hielo estimada a temperatura x = 6,573 – 1,388 x

y_{x+1}= incidencia de hielo a temperatura x+1 = 6,573 – 1,388(x +1)

$$y_{x+1} - y_x = -1{,}388$$

Con el modelo logarítmico $y_x = 5{,}993 - 3{,}382\ \log(x)$

¿Qué efecto tendrá sobre la incidencia del hielo el incrementar la temperatura un 1%? (supone incrementar 0,029ºC sobre una media de 2,9)

Respuesta: la incidencia de hielo descenderá en 0,034 meses

y_x= incidencia de hielo estimada a temperatura x = 5,993-3,382 log(x)

$Y_{1'01x}$= incidencia de hielo a temperatura 1'01x = 5,993-3,382 log(1'01x)

$$y_{1'01x} - y_x = -3{,}382 \log(1'01) = -0{,}034$$

Ejemplo 2. 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%

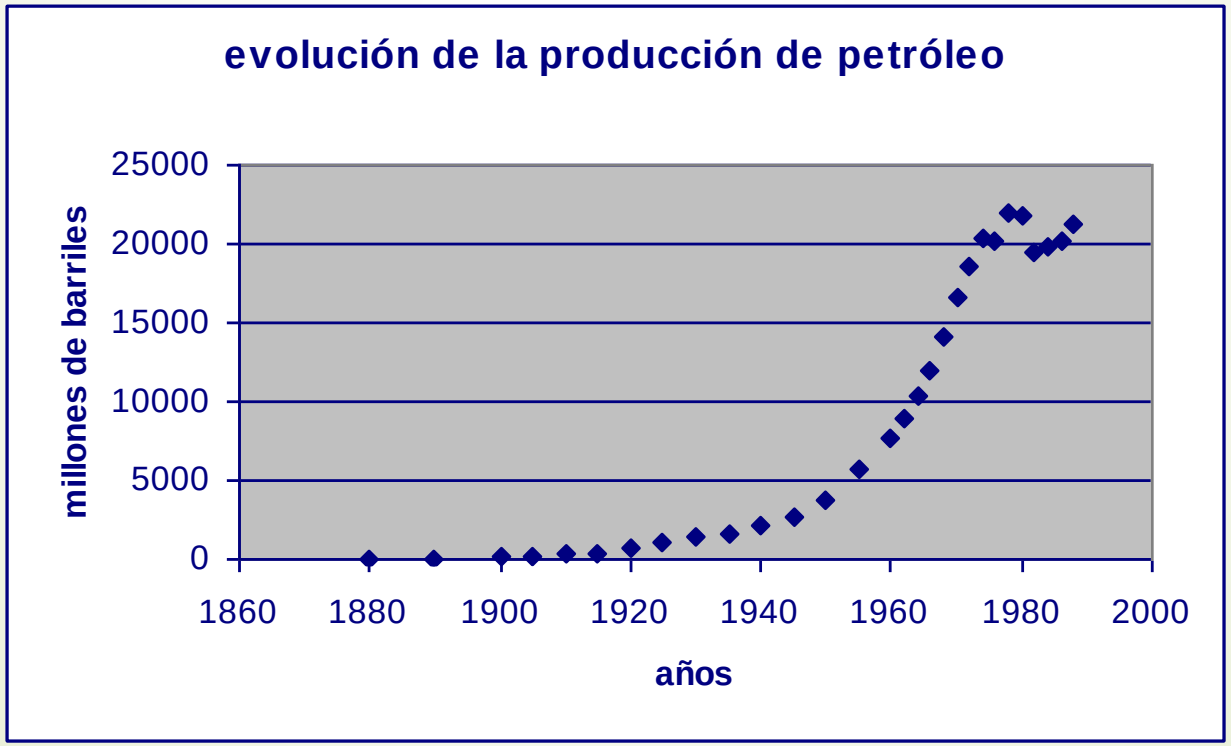
Algunos abusos que se pueden cometer en la regresión

- Extrapolación
- Generalización
- Falsa correlación
- Causalidad

Extrapolación

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

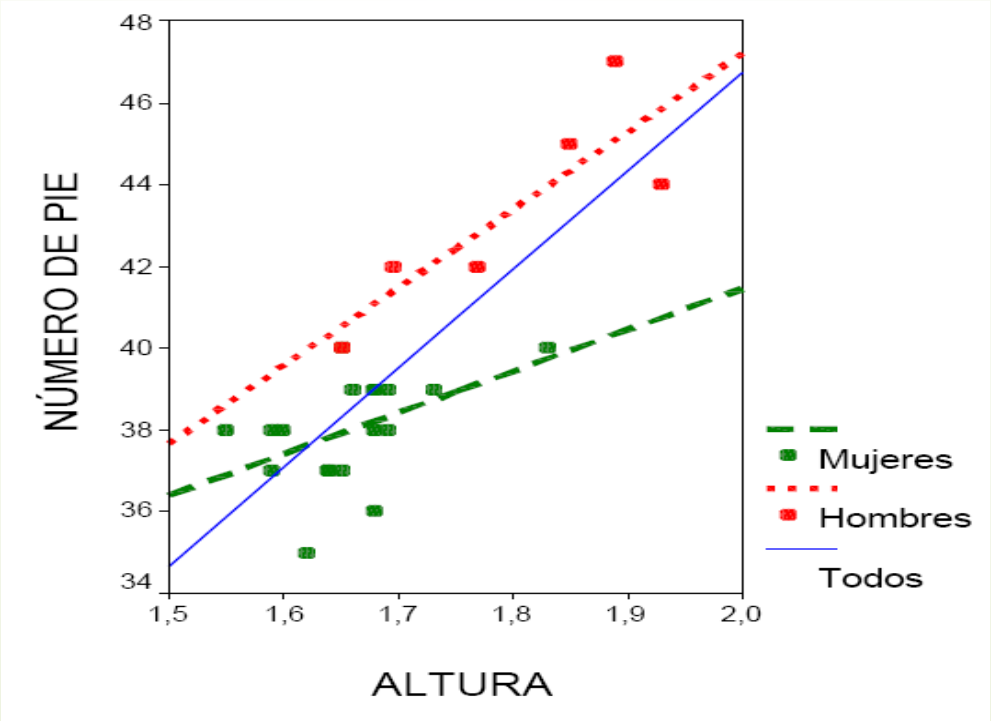
Ejemplo. Evolución de la producción de petróleo



Generalización

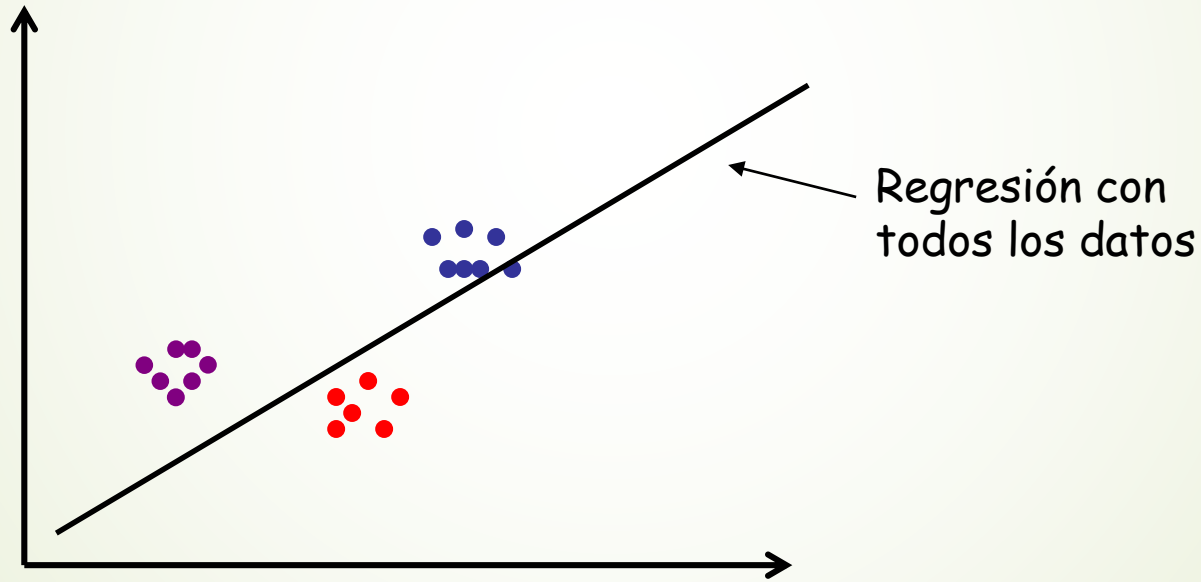
Utilizar un único modelo para conjuntos de datos que proceden de distintas poblaciones

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



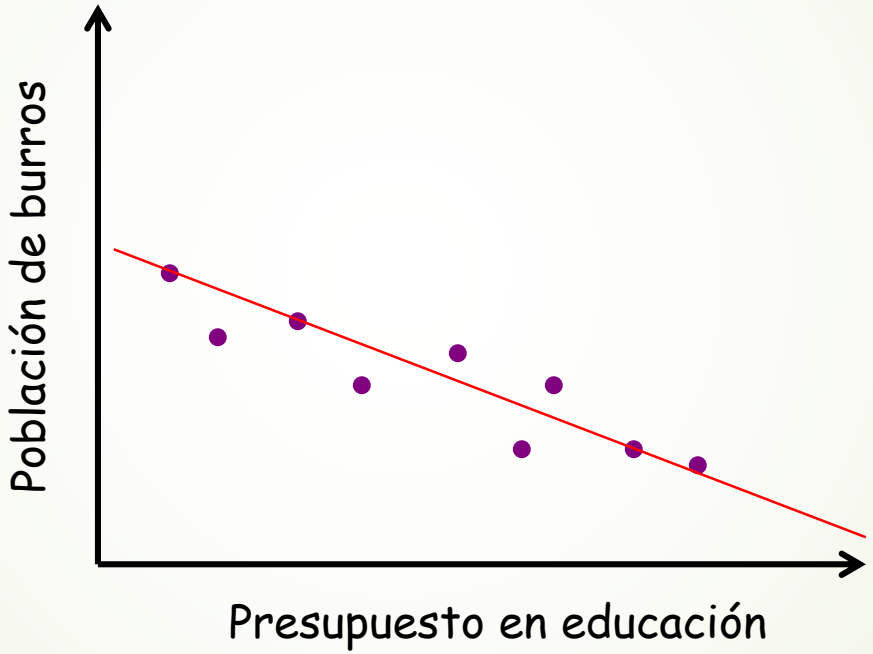
Correlación falsa

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