Análisis de datos

2º de Biología

III. Regresión lineal

Departamento de Matemáticas

Universidad Autónoma de Madrid

2011/12

El problema

Explicar la variabilidad de una magnitud continua Y —variable explicada— por medio de la variación de otra variable continua X —variable explicativa.

Ejemplos

- Peso de una persona por medio de la estatura
- Presión atmosférica en función de la altitud
- Alargamiento de un resorte en función de la fuerza aplicada

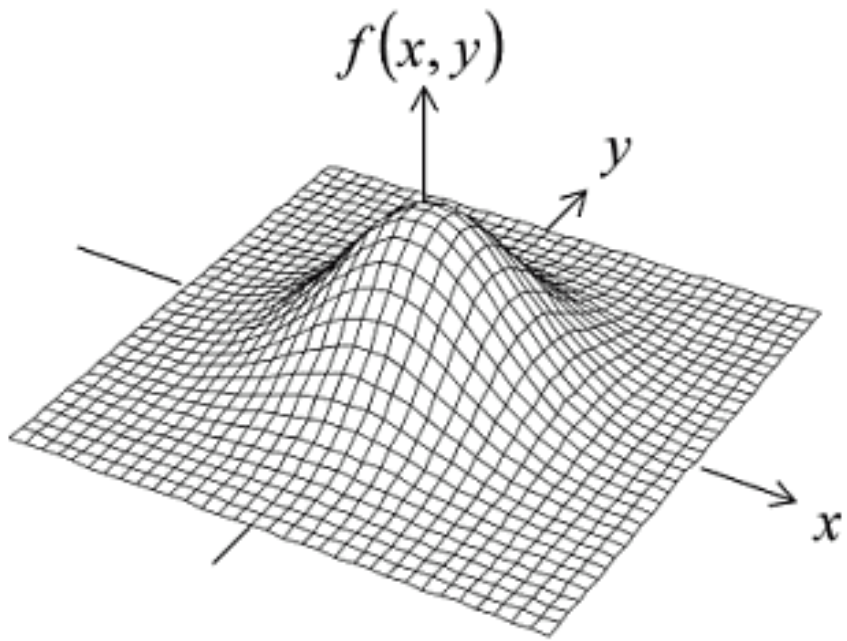
Los elementos

- Normal bivariante
- Ajuste por mínimos cuadrados

Normal bivariante

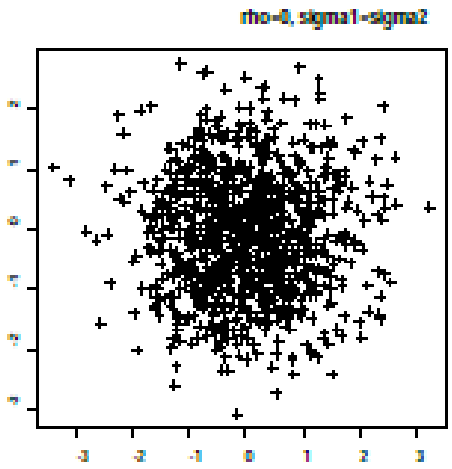
$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\left(\sigma_2^2(x-\mu_1)^2-2\sigma_1\sigma_2(x-\mu_1)(y-\mu_2)+\sigma_1^2(y-\mu_2)^2\right)\right\}$$

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

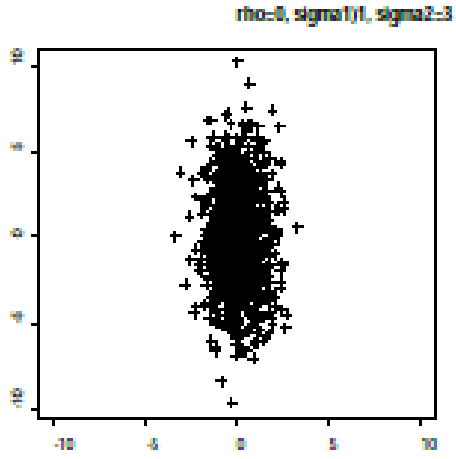


Simulación

Distribución Normal Bivariante (simulación de datos)

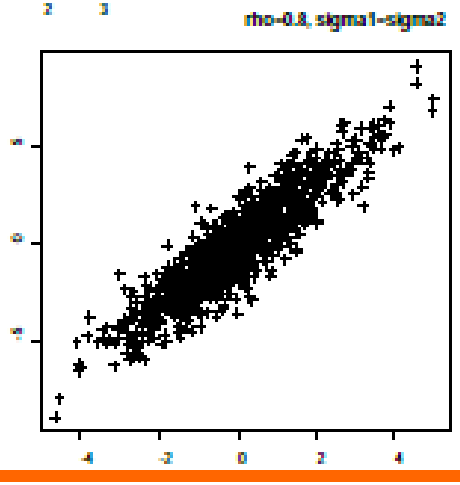


$\mu_1 = \mu_2 = 0$
$\sigma_1 = \sigma_2 = 1$
$\rho = 0$

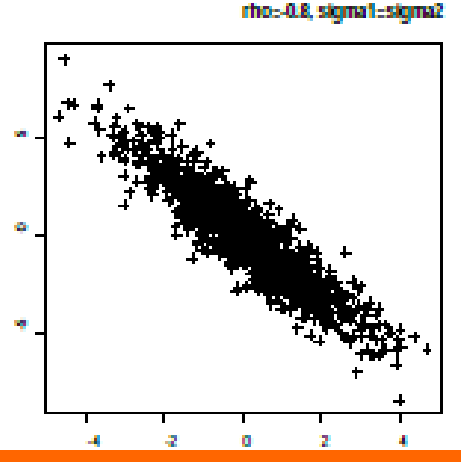


$\mu_1 = \mu_2 = 0$
$\sigma_1 = 1\ \sigma_2 = 3$
$\rho = 0$

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



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



3

Modelo

Modelo lineal: $Y = \beta_0 + \beta_1 X + U$

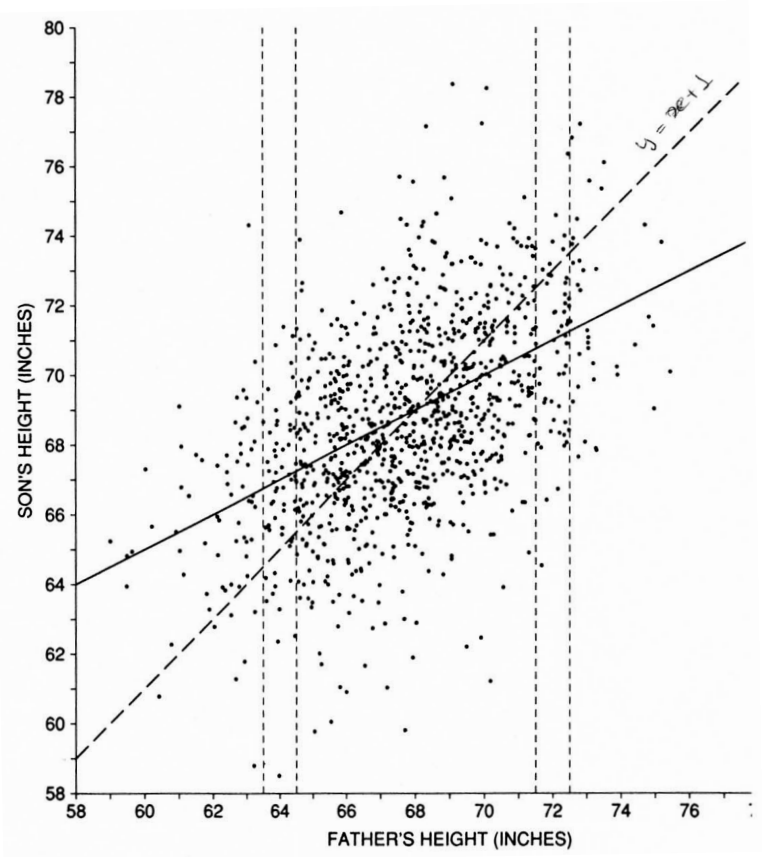
Normalidad y homocedasticidad:

$U \sim N(0, \sigma)$, es decir, $Y|_x \sim N(\beta_0 + \beta_1 x, \sigma)$

Parámetros: β_0, β_1, σ.

Ejemplo: Pearson y Lee

On the Laws of Inheritance in Man; Karl Pearson, Alice Lee; Biometrika (1903)



$Y = 0{,}516X + 33{,}73$

X: estatura del padre
Y: estatura del hijo

Datos: 1078 parejas (padre, hijo)

Estatura media padres: 68 pulgadas
Estatura media hijos: 69 pulgadas

$v_x = v_y = 2{,}7$

$r = 0{,}51$

OBSERVACIÓN: Artículo completo disponible en la página del profesor

Muestra aleatoria

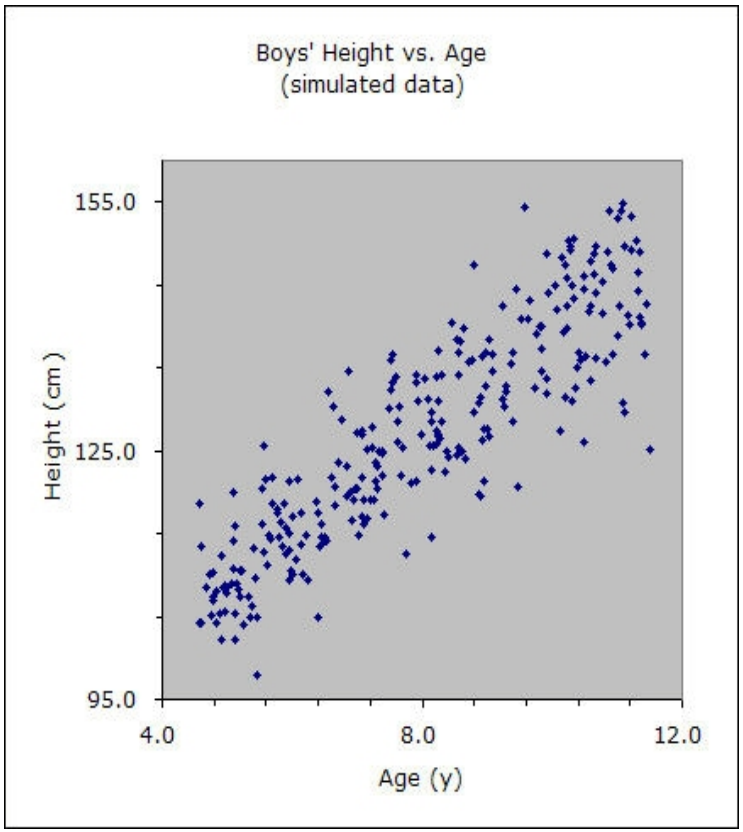
Muestra: (X_i, Y_i), $i = 1..n$, pares independientes

Dos modelos distintos:

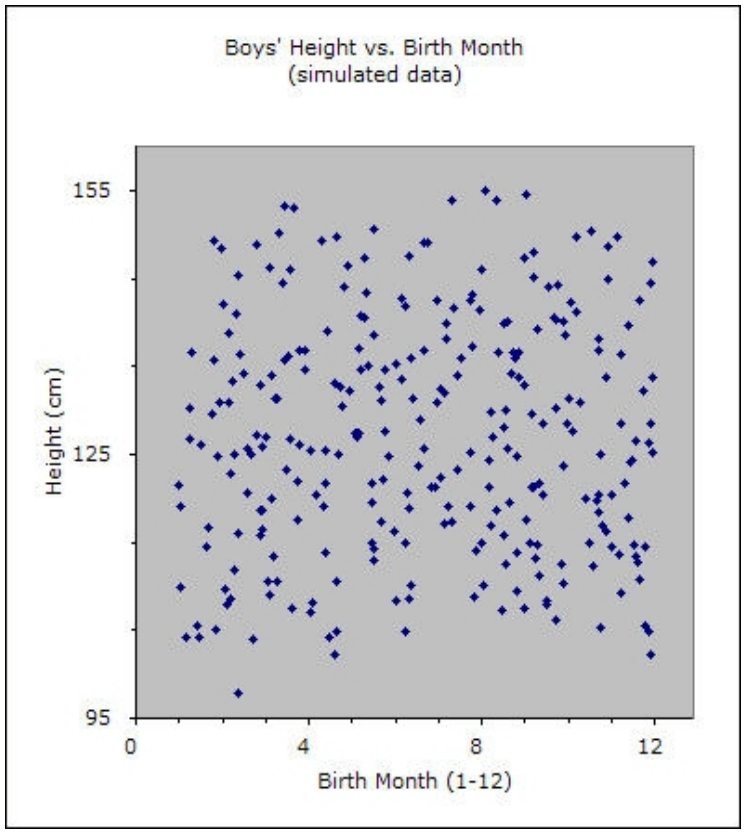
- Valor de X_i aleatorio
  - Estatura y peso de un conjunto de individuos
  - Estatura del padre y estatura del hijo para un conjunto de parejas (padre, hijo)
- Valor de X_i determinado por el investigador
  - Presión atmosférica a una serie de altitudes prefijadas
  - Extensión del resorte para una serie de masas prefijadas

Ambos modelos se tratan matemáticamente de forma análoga

Ejemplo

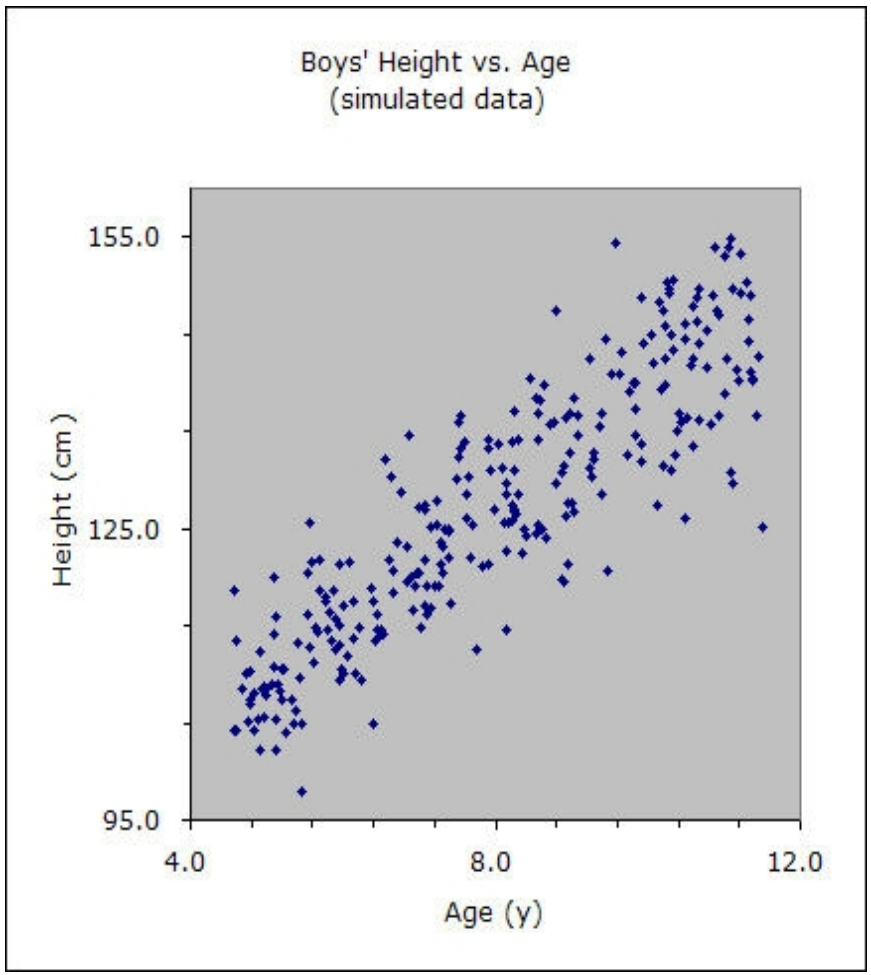


La estatura crece linealmente con la edad



No hay relación (lineal) entre el mes de nacimiento y la estatura

Ajuste por mínimos cuadrados



Mínimos cuadrados

Mínima suma de cuadrados de distancias verticales a la recta.

Determinar β_0 y β_1 para que la suma

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

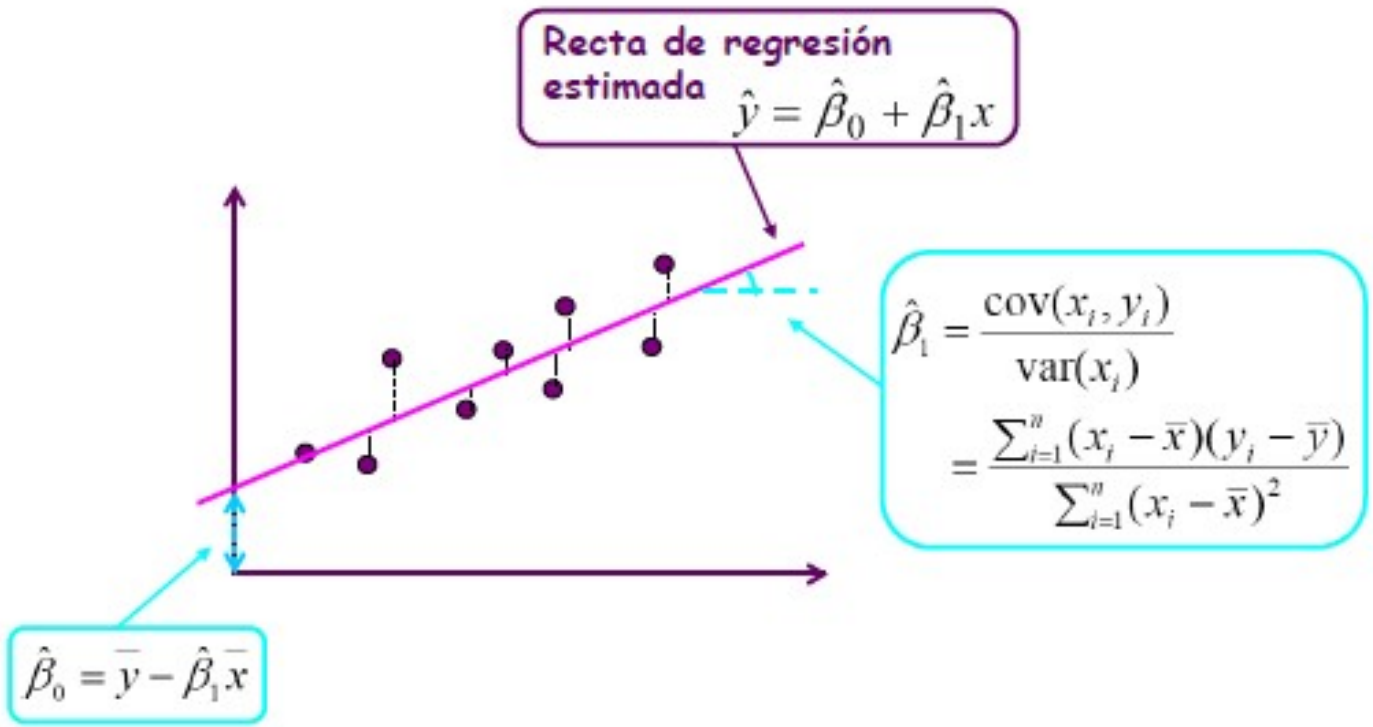
sea mínima.

Notación

$$\bar{x} = \frac{1}{n} \sum_i x_i \qquad \bar{y} = \frac{1}{n} \sum_i y_i$$

Parámetros estimados

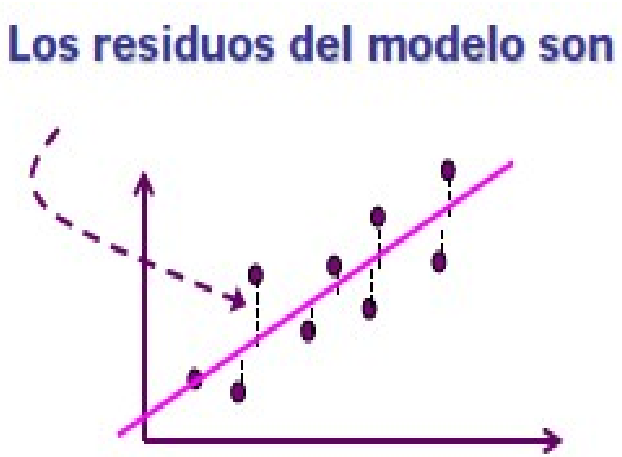
Ajuste de una recta a n pares de datos (x_i, y_i) Estimación de los coeficientes de la recta



12

Varianza residual

Estimación de σ^2:



Residuos

$$
\begin{aligned}
e_i &= y_i - \hat{y}_i \\
&= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})
\end{aligned}
$$

Varianza residual

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

Estimación puntual de los parámetros

$$\hat{\beta}_1 = \frac{cov(x,y)}{v_x}$$

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

$$\hat{\sigma}^2 = S_R^2 = \frac{\sum e_i^2}{n-2} \quad ;$$

$$\hat{\rho} = r = \frac{cov(x,y)}{\sqrt{v_x v_y}} \quad .$$

Coeficiente de determinación: $R^2 = r^2$

Recuérdese:

$$v_x = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_i x_i^2\right) - \bar{x}^2 \quad ; \quad v_y = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \left(\frac{1}{n}\sum_i y_i^2\right) - \bar{y}^2$$

$$cov(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n}\sum_i x_i y_i\right) - \bar{x}\bar{y}$$

Intervalos de confianza

$$\mathsf{IC}_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 - t_{n-2;\frac{\alpha}{2}} \cdot S_R \sqrt{\frac{1}{n v_x}}, \quad \hat{\beta}_1 + t_{n-2;\frac{\alpha}{2}} \cdot S_R \sqrt{\frac{1}{n v_x}} \right)$$

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

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

Contraste de la regresión

¿Es $\beta_1 \neq 0$?

$\mathsf{H}_0 \equiv \beta_1 = 0$

$\mathsf{H}_1 \equiv \beta_1 \neq 0$

Contraste t

$$\mathcal{R} = \left\{ \left| \frac{\hat{\beta}_1}{S_R \sqrt{\frac{1}{n v_x}}} \right| > t_{n-2;\alpha/2} \right\}$$

Contraste ANOVA

Equivalente al contraste t (los estadísticos que se obtienen con una muestra dada tiene exactamente el mismo p-valor en ambos contrastes)

Sumas de cuadrados

$$\text{SCT} = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 = \text{SCR} + \text{SCE}$$

Tabla ANOVA

	Suma de cuadrados	g. de l.	Varianzas	F
Modelo	SCE	1	SCE	SCE/S_R^2
Error	SCR	$n-2$	S_R^2	
Total	SCT	$n-1$		

$$\mathcal{R} = \{F > F_{1,n-2;\alpha}\}$$

Comentarios

- El contraste de la regresión permite decidir si parte de la variabilidad de la Y puede atribuirse a la X.
- Las sumas de cuadrados pueden calcularse a partir de algunos estadísticos ya calculados:

  $\text{SCT} = nv_y = (n-1)s_y^2$

  $\text{SCE} = nv_y r^2 = (n-1)s_y^2 r^2$

  $\text{SCR} = nv_y(1-r^2) = (n-1)s_y^2(1-r^2)$
- Según lo anterior:

$$F = (n-2)\frac{r^2}{1-r^2}$$

Coeficiente de determinación: R^2

- Da una idea de qué fracción de la variablidad de Y está explicada por X.
- Su valor, $R^2 = \frac{SCE}{SCT}$, es siempre positivo.
- En regresión lineal simple coincide con r^2, el cuadrado del coeficiente de correlación.

Ejemplo: *Oecanthus niveus*; Bessey & Bessey; The American Naturalist; 1897

T	N
61	103
67	123
72	150
75	171
81	190

$$n = 5$$

$$\sum \mathsf{T} = 356 \qquad \sum \mathsf{T}^2 = 25\,580$$

$$\sum \mathsf{N} = 737 \qquad \sum \mathsf{N}^2 = 113\,579$$

$$\sum \mathsf{NT} = 53\,539$$

Modelo para explicar la temperatura (T) por medio del número de «cricks» por minuto (N): $\mathsf{T} = \beta_0 + \beta_1 \mathsf{N}$

(...)

$$\bar{\mathsf{T}} = 71{,}20 \qquad \bar{\mathsf{T}}^2 = 5116{,}00$$
$$\bar{\mathsf{N}} = 147{,}40 \qquad \bar{\mathsf{N}}^2 = 27\,715{,}80$$
$$nv_{\mathsf{T}} = (n-1)s_{\mathsf{T}}^2 = 232{,}80$$
$$nv_{\mathsf{N}} = (n-1)s_{\mathsf{N}}^2 = 4945{,}20 \qquad n\mathrm{cov}(\mathsf{N},\mathsf{T}) = 1064{,}60$$

$$\hat{\beta}_1 = \frac{\mathrm{cov}(\mathsf{N},\mathsf{T})}{v_{\mathsf{N}}} = \frac{n\mathrm{cov}(\mathsf{N},\mathsf{T})}{nv_{\mathsf{N}}} = 0{,}2153$$

$$\hat{\beta}_0 = \bar{\mathsf{T}} - \hat{\beta}_1\bar{\mathsf{N}} = 39{,}47$$

Ecuación de la recta de regresión: $\mathsf{T} = 39{,}47 + 0{,}215\mathsf{N}$

$$r = \frac{\mathrm{cov}(\mathsf{N},\mathsf{T})}{\sqrt{v_{\mathsf{N}}v_{\mathsf{T}}}} = 0{,}9922$$

$$S_R^2 = \frac{nv_{\mathsf{T}}}{n-2}(1-r^2) = 1{,}206$$

(...)

Intervalos de confianza:

$\text{IC}_{95\,\%}(\beta_0) = (31{,}98\,,\ 46{,}96)$

$\text{IC}_{95\,\%}(\beta_1) = (0{,}166\,,\ 0{,}265)$

Sumas de cuadrados:

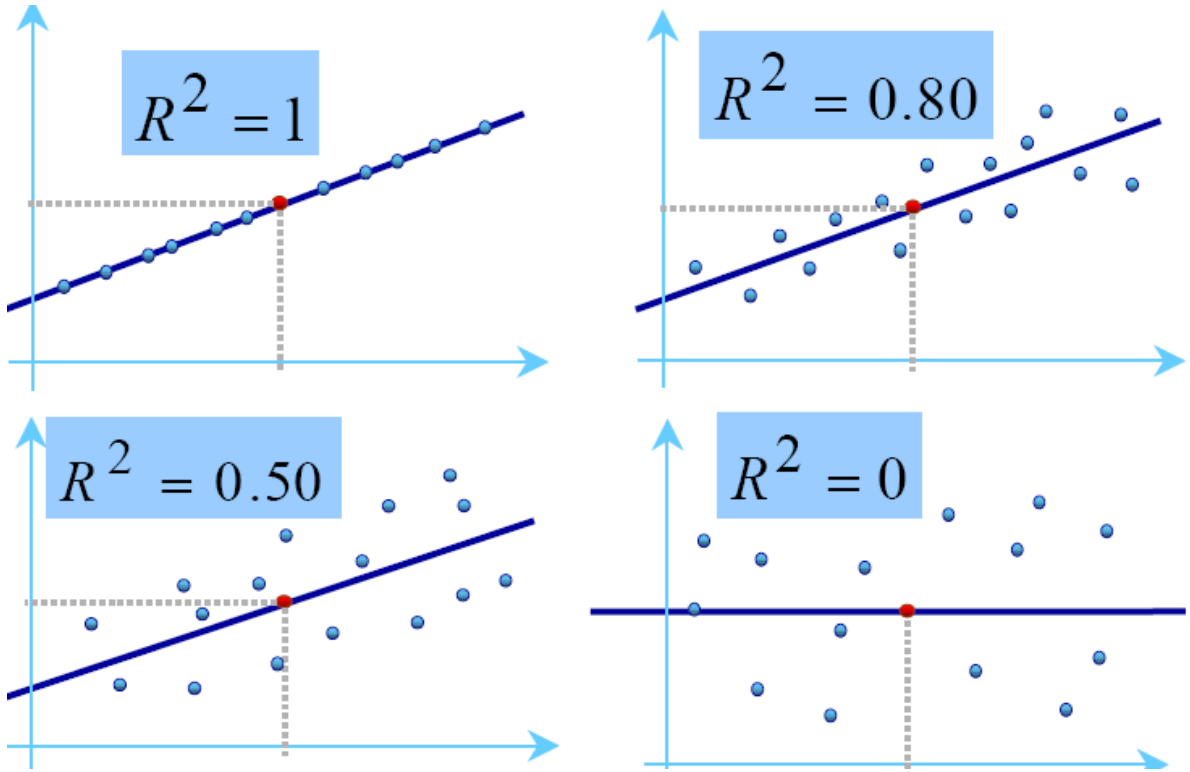
$\text{SCE} = 229{,}2$ $\qquad$ $\text{SCR} = 3{,}618$ $\qquad$ $\text{SCT} = 232{,}8$

Contraste F: $\qquad$ $F = 190$ $\qquad$ $F_{1,3;0{,}05} = 10{,}13$

Al nivel de significación $\alpha = 0{,}05$, se rechaza $\beta_1 = 0$

Coeficiente de determinación: $\qquad$ $R^2 = 0{,}985$

Interpretación



Observación de los datos

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/

Observación de los datos

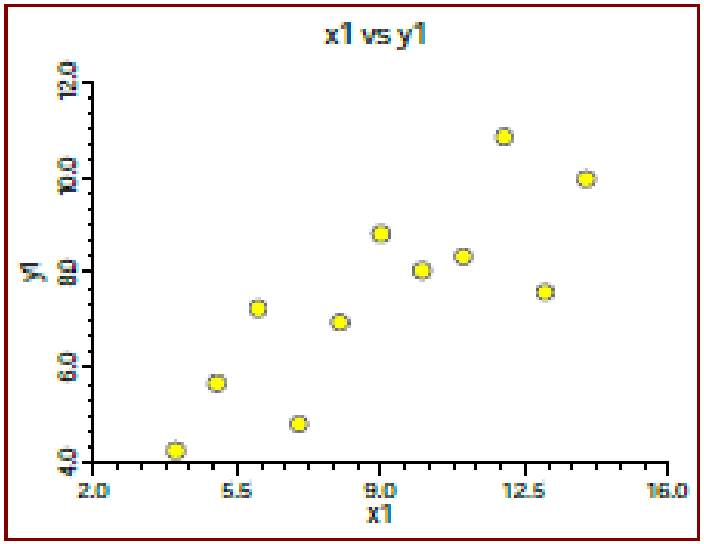
Los 4 grupos de datos tienen exactamente los mismos valores descriptivos siguientes:

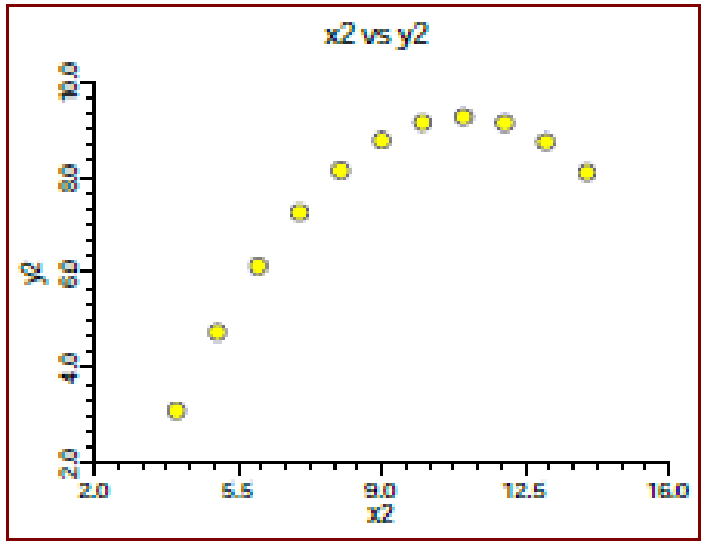
Número de datos	11
Media de las x's	9.0
Media de las y's	7.5
Ecuación de la recta de regresión	$y = 3 + 0.5x$

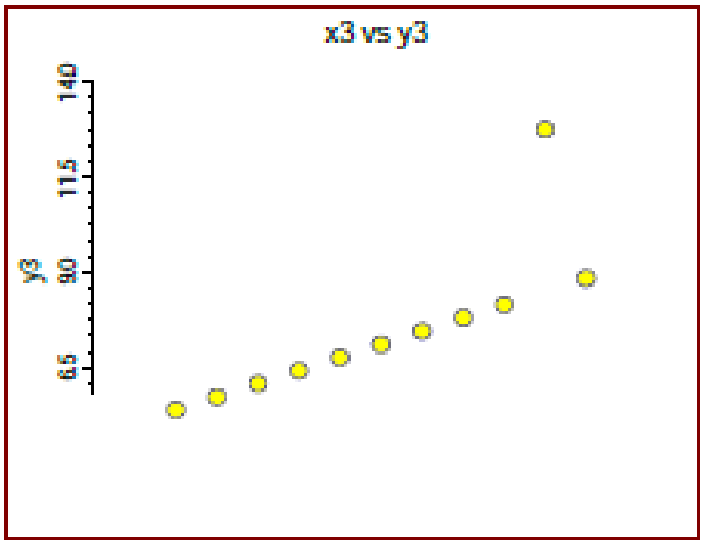
Coeficiente de correlación	0.82
r^2	0.67

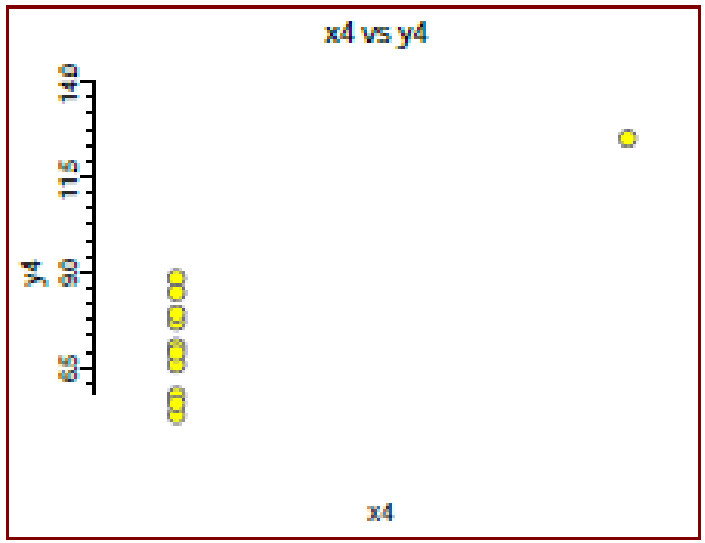
Observación de los datos

Pero los gráficos son:









Requisitos

- Linealidad
- Normalidad
- Homocedasticidad
- Independencia

Desviaciones significativas sobre estos requisitos pueden proporcionar conclusiones incorrectas

Análisis de residuos

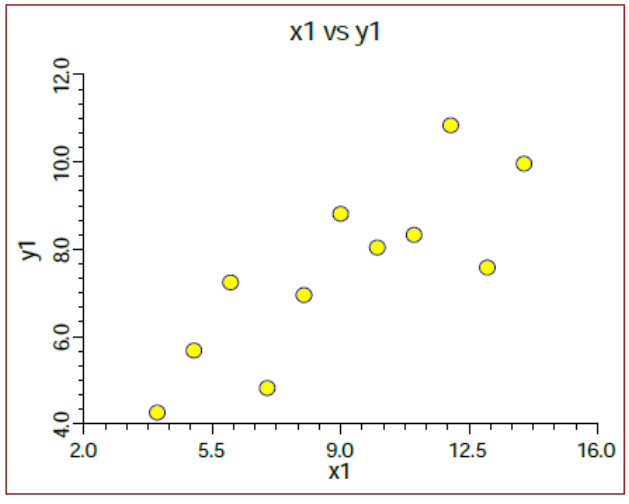
Residuo: $e_i = y_i - \hat{y}_i$

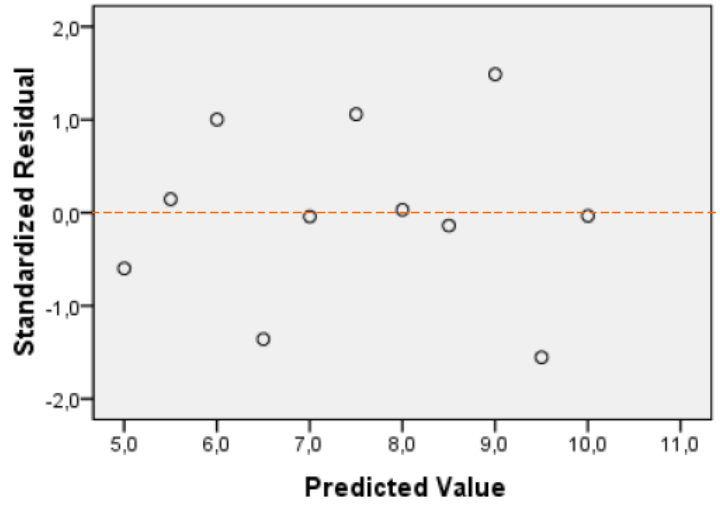
Observación: $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$

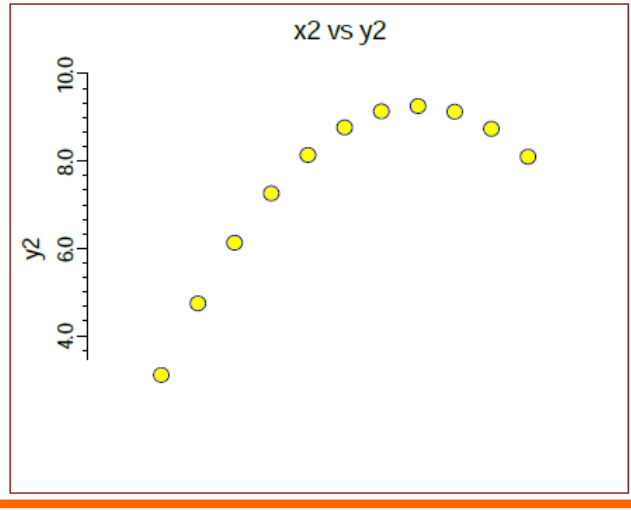
Representación gráfica de los residuos

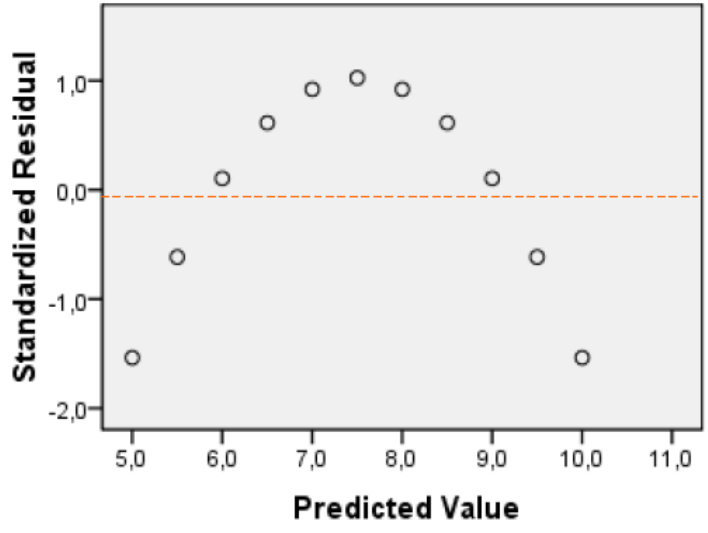
Debe hacerse respecto de los valores pronosticados $(\hat{y}_i)$, nunca respecto de los valores observados (y_i) de la variable explicada.

Los cuatro ejemplos con los mismos descriptivos

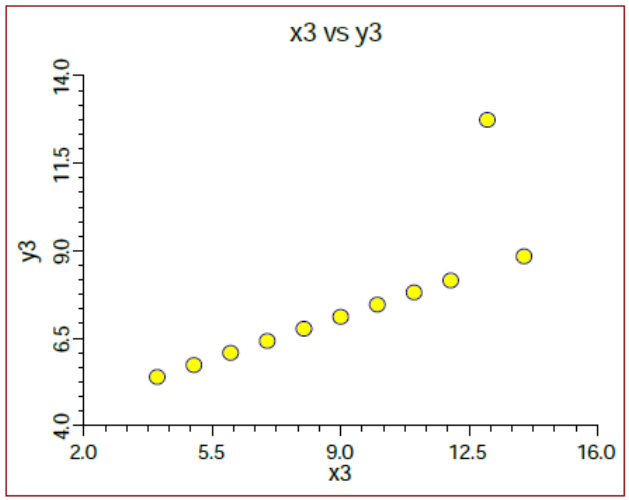


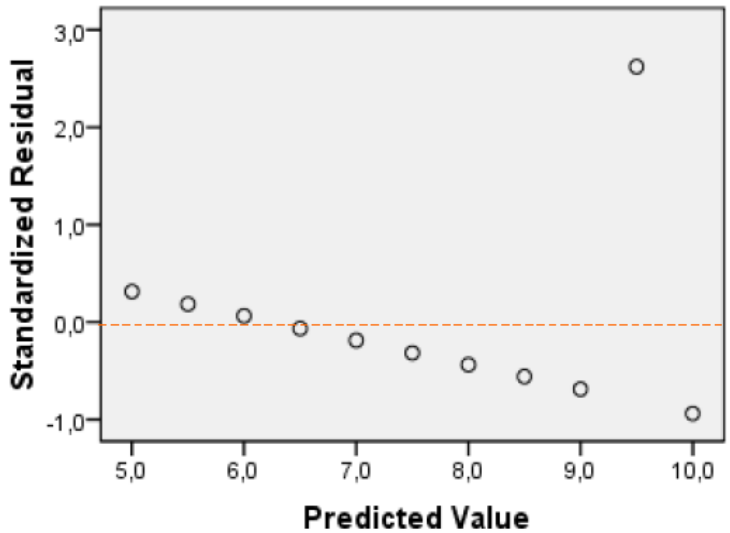


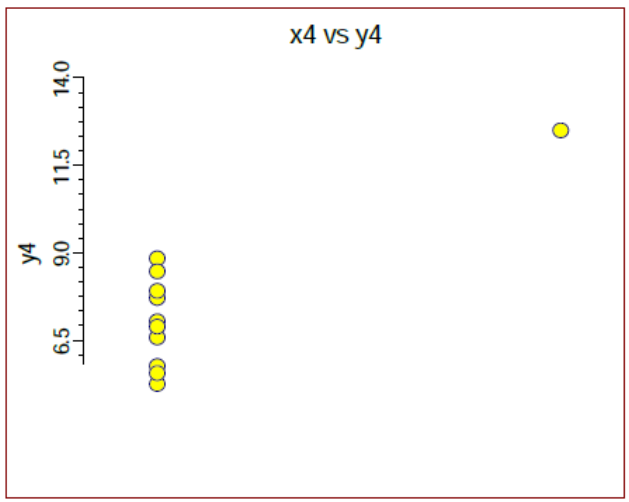


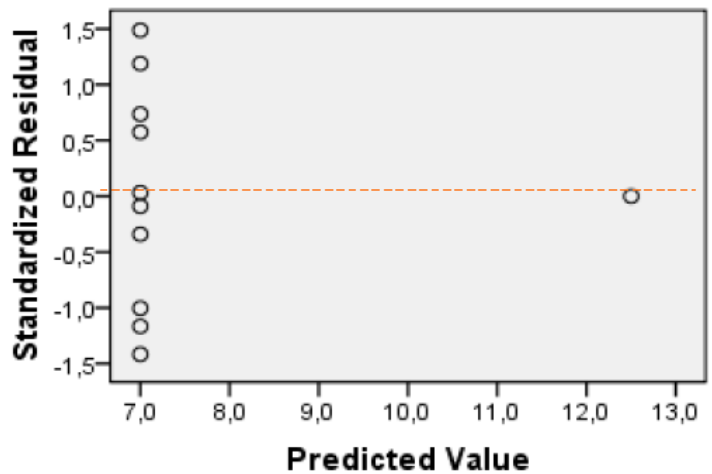


Casos 3 y 4


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

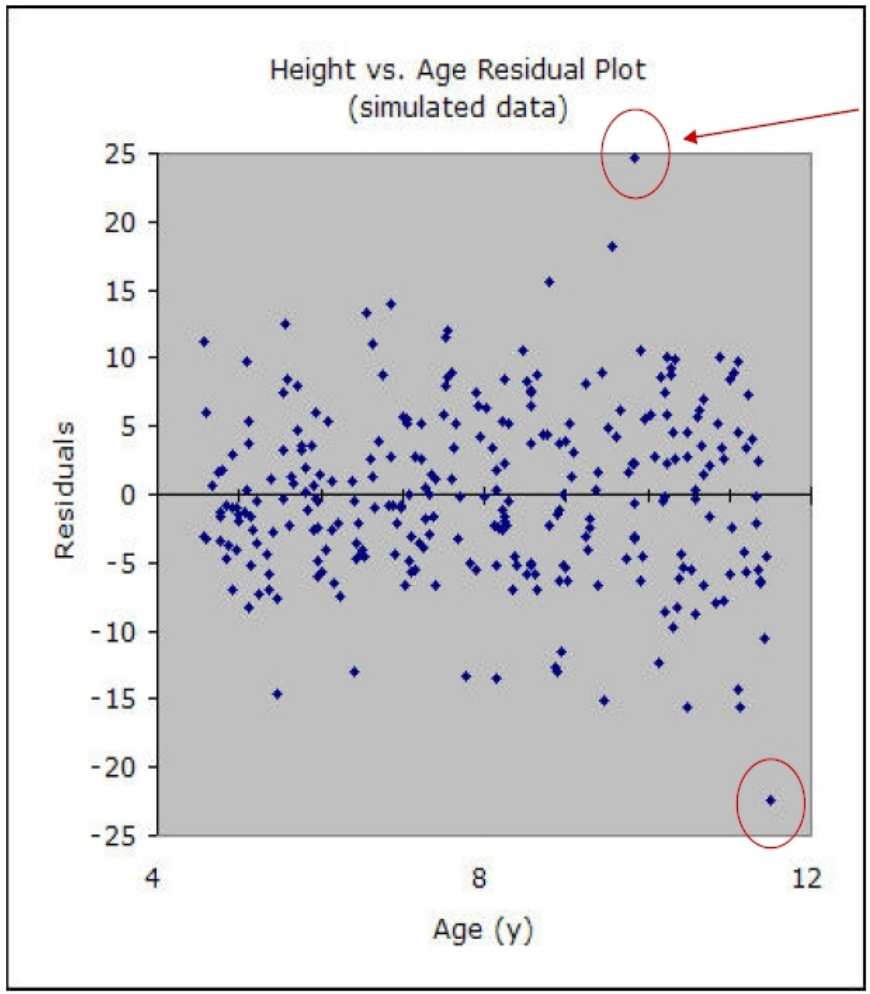

Standardized Residual
3,0
2,0
1,0
0,0
-1,0
5,0
6,0
7,0
8,0
9,0
10,0
Predicted Value


x4 vs y4
y4
14.0
11.5
9.0
6.5


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

Estatura vs. edad

Gráfico de los residuos e_i



¿es este un valor anómalo?

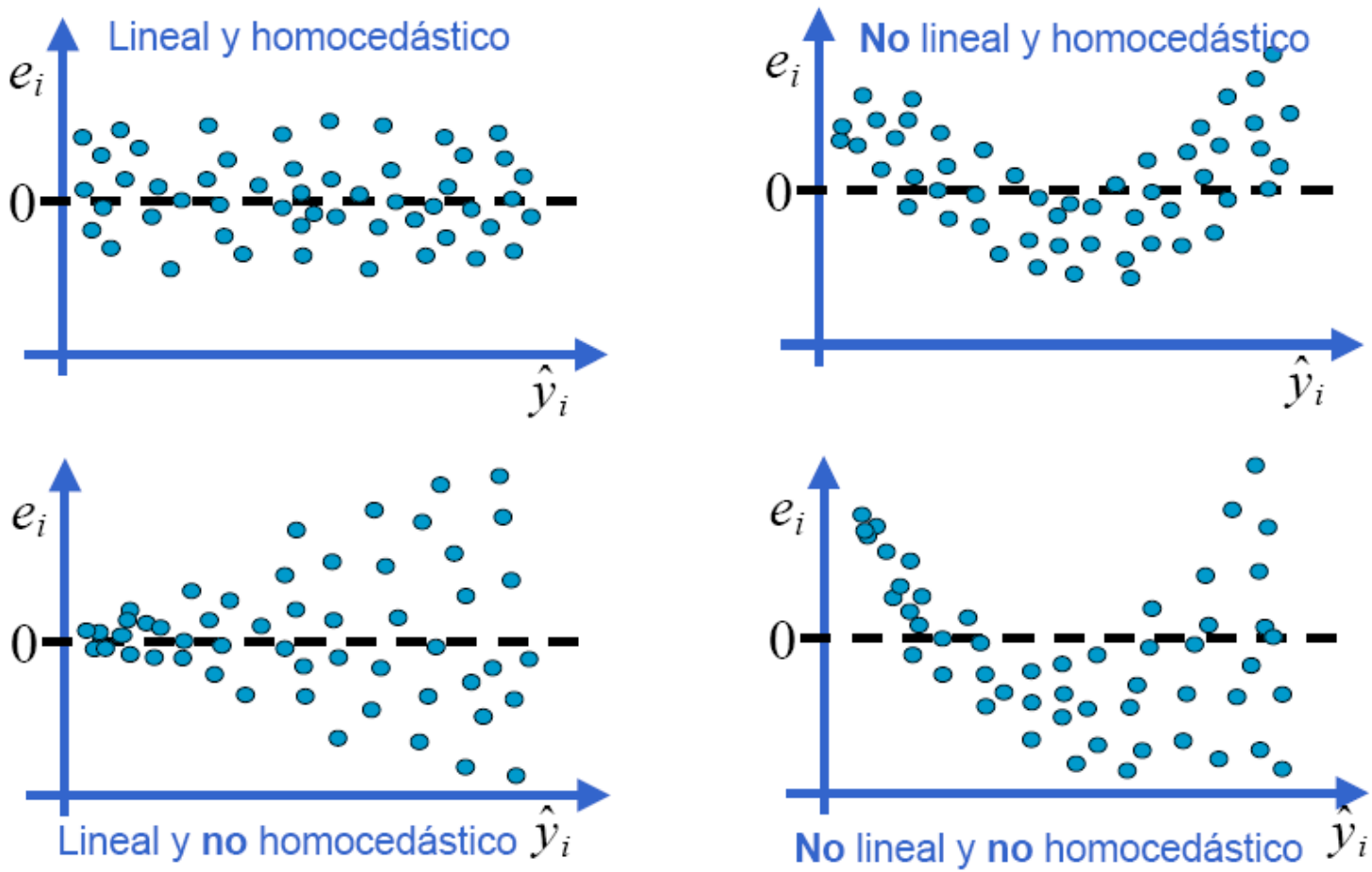
$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

En abcisas los valores de x_i (edades en años)

En ordenadas los residuos e_i sin tipificar

Dispersión de residuos

Las hipótesis del modelo

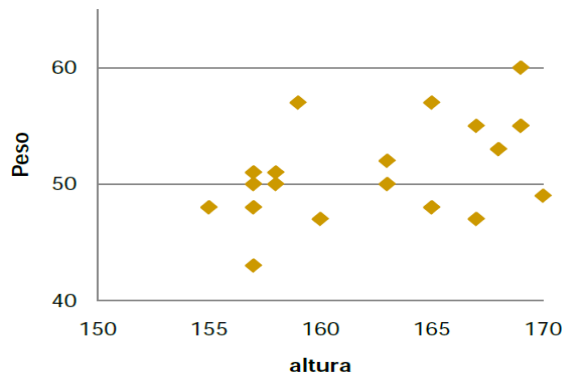


Ejemplo

La siguiente tabla recoge los datos de altura (cm) y peso (kg) de 20 mujeres estudiantes de la UAM

estatura	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

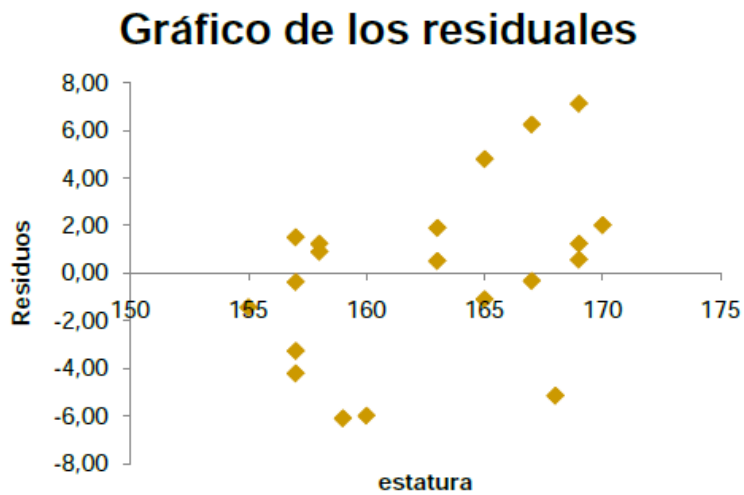


Coeficiente de correlación:
$= 0{,}476$

Estimaciones:
$\beta_0 = -11{,}84$
$\beta_1 = 0{,}388$

Residuos

Dato nº	Pronóstico	Residuo
1	49,88	7,12
2	50,27	−3,27
3	53,38	−0,38
4	49,11	0,89
5	49,11	−6,11
6	48,33	−0,33
7	52,21	−4,21
8	49,11	−1,11
9	52,99	2,01
10	51,44	0,56
11	53,77	1,23
12	49,49	0,51
13	53,77	6,23
14	49,49	1,51
15	49,11	1,89
16	51,44	−1,44
17	54,15	−5,15
18	52,21	4,79
19	52,99	−5,99
20	53,77	1,23



Tranformaciones de datos

Cuando detectamos problemas de no linealidad o heterocedasticidad y queremos aplicar las técnicas de regresión lineal

Ejemplos

Logaritmo:

$$y = ke^{\beta x} \longrightarrow \ln y = \ln k + \beta x$$

Doble logaritmo:

$$y = kx^{\beta} \longrightarrow \log y = \log k + \beta \log x$$

NOTA: Cualquier logaritmo

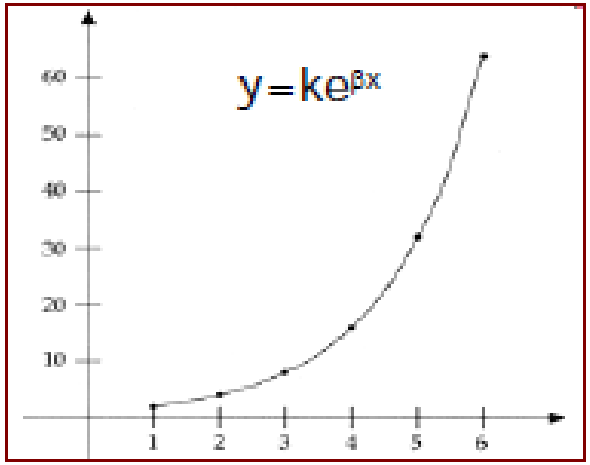
Inversa:

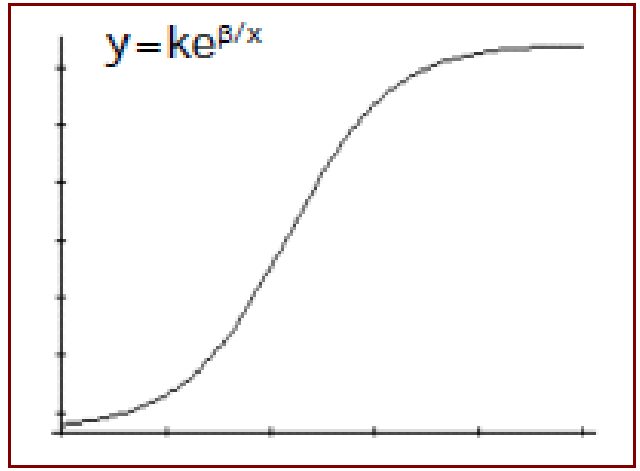
$$y = k + \beta \frac{1}{x}$$

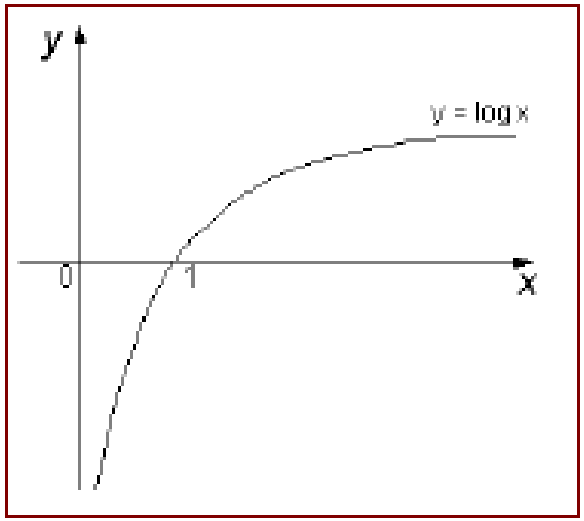
Logaritmo e inversa:

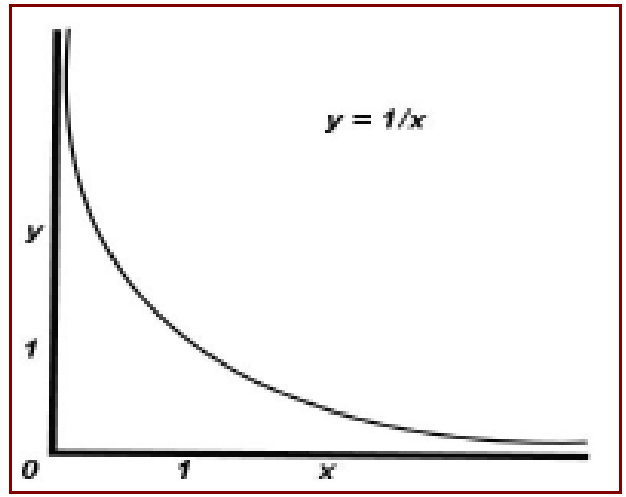
$$y = ke^{\frac{\beta}{x}} \longrightarrow \ln y = \ln k + \beta \frac{1}{x}$$

Algunas gráficas

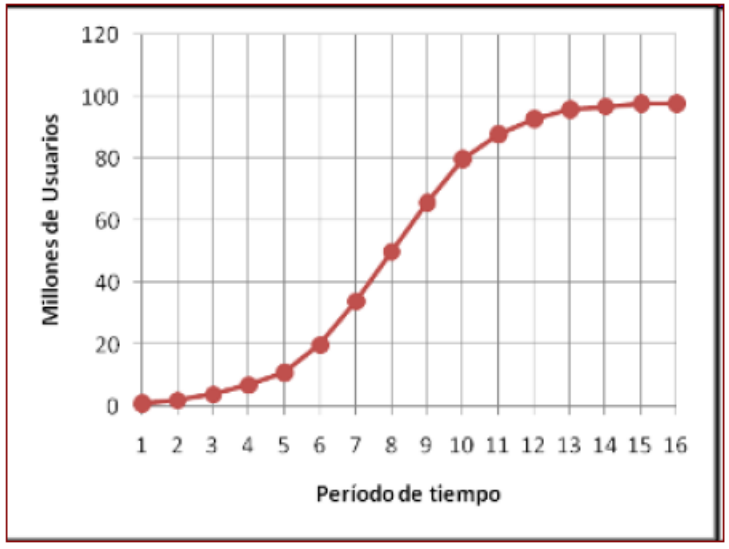








La curva logística



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

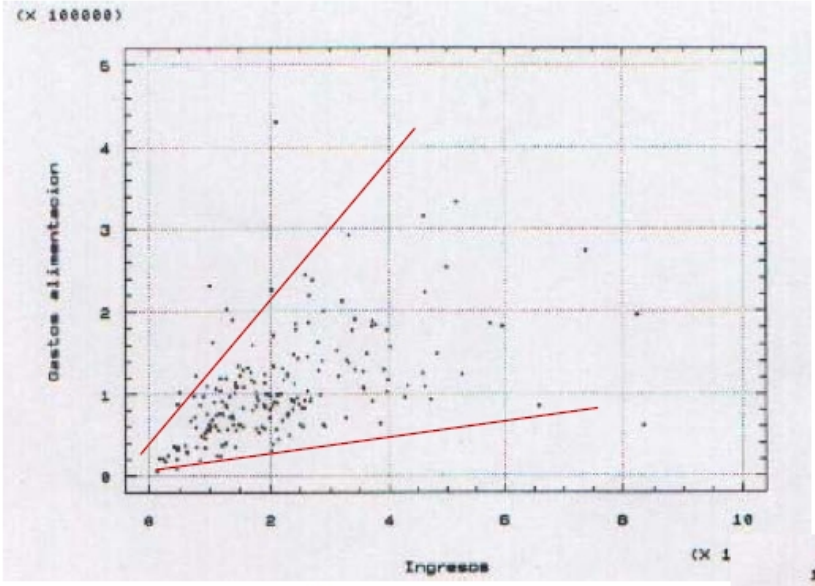
Nota: C es el valor máximo posible de la variable Y

Cambio de variable:

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

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

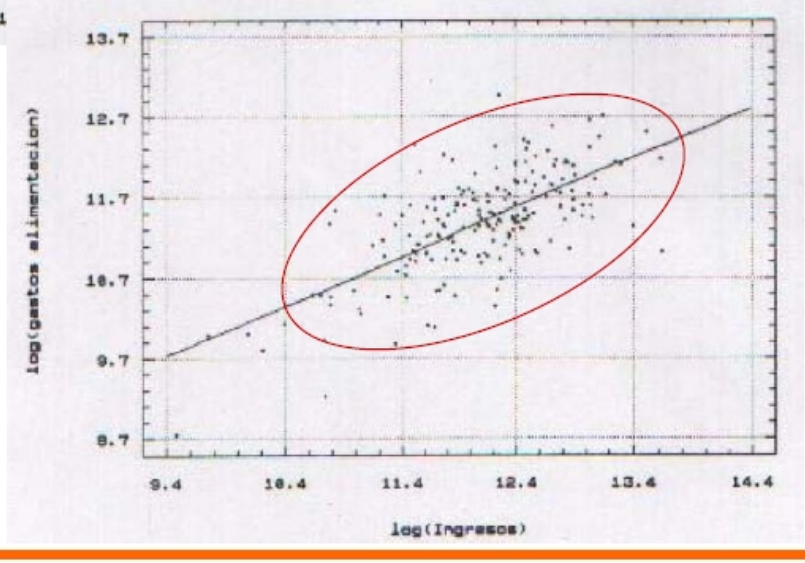
Problemas de Heterocedasticidad



Transformamos

Transformación doble log

Transformamos



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 dado por la recta de regresión ajustada: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
¿Cuál es la diferencia?

Intervalos de predicción

Estimación de la media

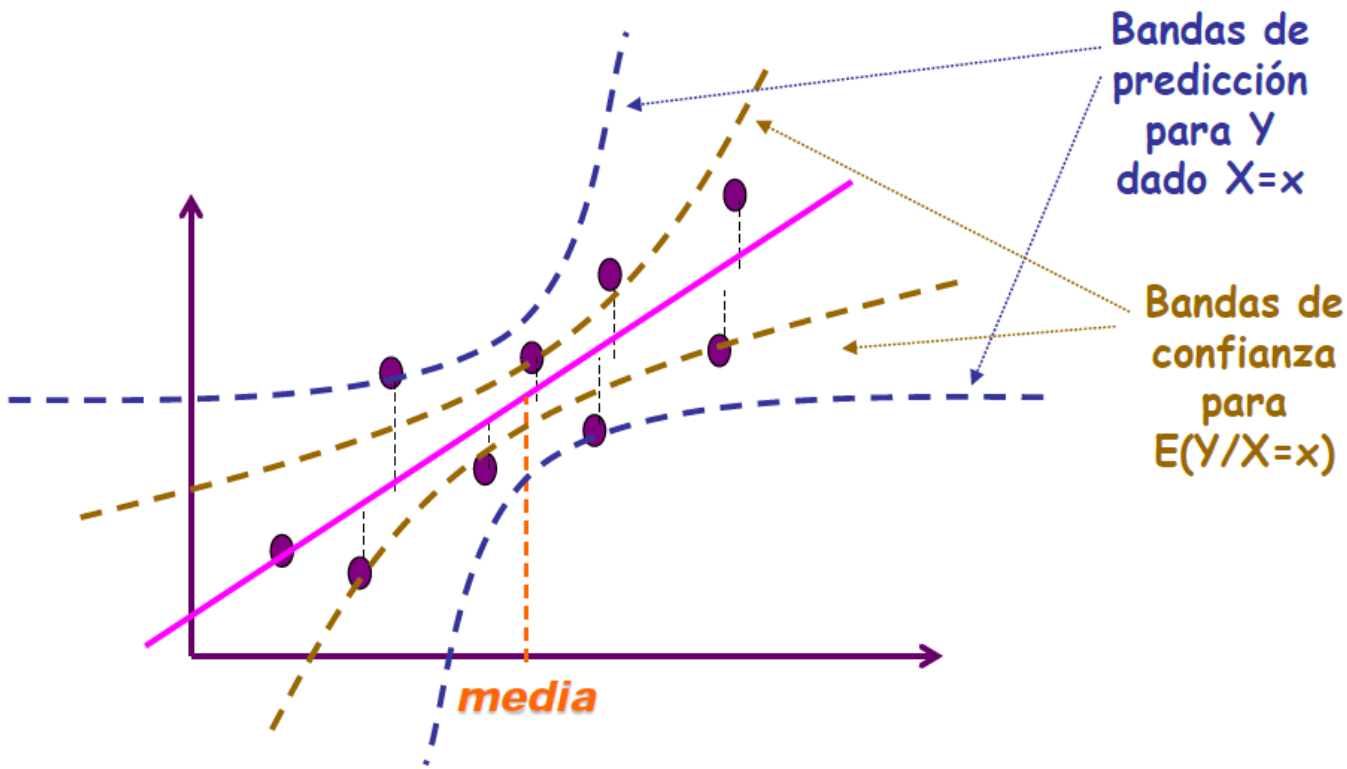
$$\mathsf{IC}_{1-\alpha}(E(Y|_{X=x_0}) = \left(\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2;\alpha/2} S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n v_x}}\right)$$

Predicción

$$\mathsf{I}_{1-\alpha}(Y|_{X=x_0}) = \left(\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2;\alpha/2} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n v_x}}\right)$$

Bandas

Los intervalos anteriores definen dos bandas en torno a la recta de regresión que tienen la misma forma. La banda para la media es siempre más estrecha que la banda para la predicción.



Formulario

Modelo: $Y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2)$ independientes, $i = 1, ..., n$.

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

$$\hat{\beta}_0 = \bar{y} - \frac{\text{cov}}{v_x}\bar{x} = \bar{y} - \hat{\beta}_1\bar{x}$$

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

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

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

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

Formulario

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

$$SCE = n v_y r^2 \ ; \qquad SCR = n v_y (1 - r^2) \ ; \qquad \text{donde } r = \frac{\text{cov}}{\sqrt{v_x v_y}}$$

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

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