TEMA 3
Modelo de regresión simple

José R. Berrendero
Departamento de Matemáticas
Universidad Autónoma de Madrid

Análisis de Datos - Grado en Biología

Estructura de este tema

- Planteamiento del problema. Ejemplos.
- El modelo de regresión lineal simple.
- Recta de regresión de mínimos cuadrados.
- Estimación, IC y contrastes para los parámetros del modelo.
- Análisis de la varianza en el modelo de regresión lineal simple.
- Predicción.
- Algunos modelos linealizables.
- Diagnóstico del modelo.

Ejemplo: temperatura y vibración de las alas

Los grillos son ectotermos, por lo que sus procesos fisiológicos y su metabolismo están influidos por la temperatura. Con el fin de estudiar estas cuestiones se ha medido el número de vibraciones por segundo de las alas de un grupo de grillos a varias temperaturas.

Vibraciones/seg.	Temp.
20.0	88.6
16.0	71.6
19.8	93.3
18.4	84.3
17.1	80.6
15.5	75.2
14.7	69.7
17.1	82.0
15.4	69.4
16.2	83.3
15.0	78.6
17.2	82.6
16.0	80.6
17.0	83.5
14.1	76.3

Ejemplo: Temperatura y vibración de las alas

Consideramos dos variables (fichero `grillos.sav`):

- X: Temperatura
- Y: Número de vibraciones de las alas por segundo

¿Qué podemos decir sobre la relación entre las dos variables?

¿Podemos afirmar (con un nivel de significación dado) que al aumentar la temperatura, aumenta la frecuencia de vibración?

¿Podemos predecir aproximadamente el valor de la variable Y si sabemos el valor de X? ¿Qué grado de fiabilidad tiene la predicción?

Ejemplo: renta y fracaso escolar en la CAM

EL PAÍS, martes 18 de octubre de 2005

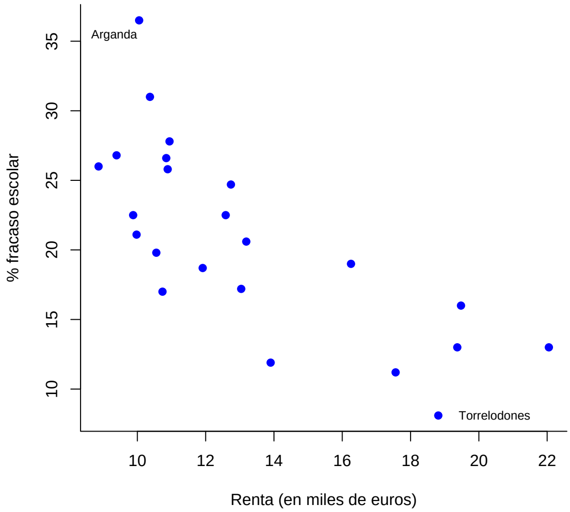
El fracaso escolar es más alto en las zonas con menor renta

Fracaso escolar en la Comunidad de Madrid

Renta per capita bruta media en 2003: 13.095 euros

	CURSO 2003/2004	
	Renta (euros)	Fracaso escolar (%)
Parla	8.864	26,0
Fuenlabrada	9.391	26,8
Leganés	9.877	22,5
Móstoles	9.977	21,1
Arganda	10.052	36,5
Torrejón	10.369	31,0
Getafe	10.555	19,8
Coslada	10.736	17,0
Pinto	10.846	26,6
Alcorcón	10.888	25,8
Alcalá de Henarés	10.942	27,8
Collado	11.913	18,7
Colmenar Viejo	12.587	22,5
Arroyomolinos	12.740	24,7
S. Sebastián de los Reyes	13.041	17,2
S. Lorenzo del Escorial	13.189	20,6
Rivas	13.903	11,9
Alcobendas	16.256	19,0
Tres Cantos	17.562	11,2
Torrelodones	18.812	8,1
Boadilla	19.368	13,0
Majadahonda	19.477	16,0
Pozuelo	22.050	13,0

EL PAÍS



Arganda
Torrelodones
% fracaso escolar
10
15
20
25
30
35
10
12
14
16
18
20
22
Renta (en miles de euros)

Covarianza

Se dispone de un conjunto de n pares de observaciones

$$(x_1, y_1), \ldots, (x_n, y_n).$$

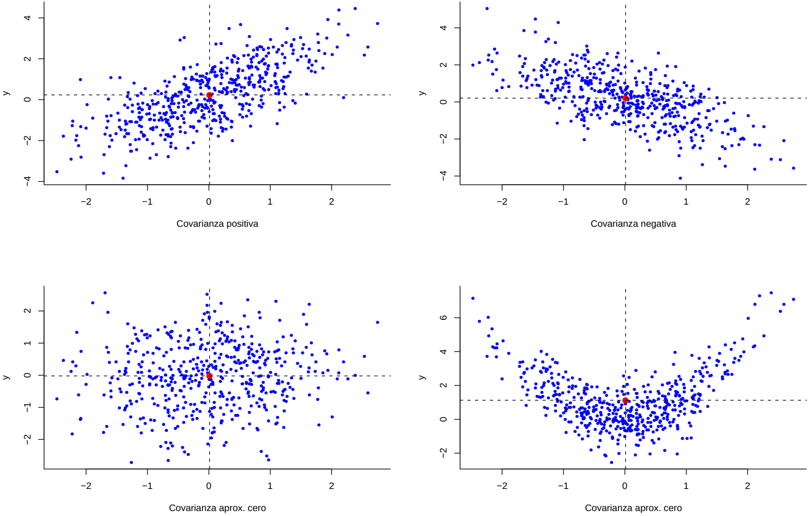
La covarianza entre x e y sirve para cuantificar el grado de relación lineal que hay entre x e y:

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

Propiedades:

- $\mathrm{cov}_{xy} = \mathrm{cov}_{yx}$.
- cov_{xy} depende de las unidades en que se miden x e y.
- $\mathrm{cov}_{xx} = \mathrm{v}_x$, es decir, la covarianza de x con x es la varianza de x.

Interpretación de la covarianza



Coeficiente de correlación

Resulta conveniente disponer de una medida de relación lineal que no dependa de las unidades. Para ello, se normaliza cov_{xy} dividiendo por el producto de desviaciones típicas, lo que lleva al **coeficiente de correlación**:

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{\sqrt{v_x}\sqrt{v_y}}.$$

Propiedades:

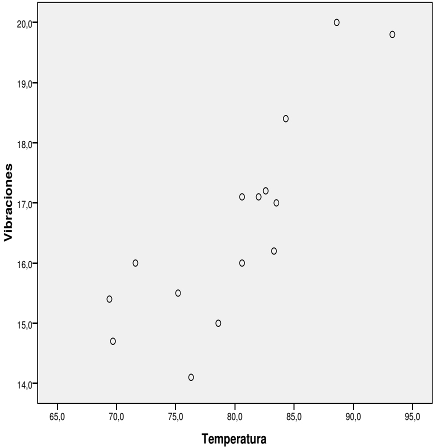
- No depende de las unidades
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Estadísticos descriptivos

	Media	Desviación típica	N
Vibraciones	16,633	1,7319	15
Temperatura	79,973	6,7170	15

Correlaciones

		Vibraciones	Temperatura
Vibraciones	Correlación de Pearson	1	,836
	Sig. (bilateral)		,000
	N	15	15
Temperatura	Correlación de Pearson	,836	1
	Sig. (bilateral)	,000	
	N	15	15



Problema de regresión

Observamos dos variables, X e Y, el objetivo es analizar la relación existente entre ambas de forma que podamos predecir o aproximar el valor de la variable Y a partir del valor de la variable X.

- La variable Y se llama **variable respuesta**
- La variable X se llama **variable regresora o explicativa**

En un problema de regresión (a diferencia de cuando calculamos el coeficiente de correlación) el papel de las dos variables no es simétrico.

Recta de regresión

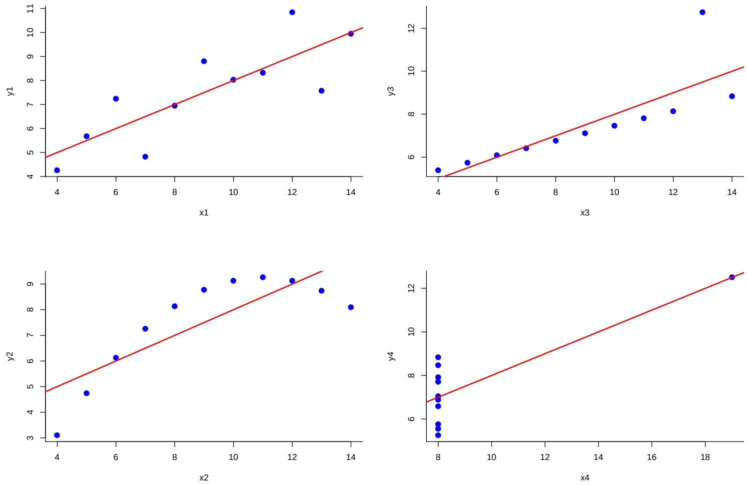
Frecuentemente, existe entre las variables una relación aproximadamente lineal:

$$Y_i \approx \beta_0 + \beta_1 x_i.$$

- La recta $y = \beta_0 + \beta_1 x$ es una **recta de regresión**.
- El parámetro β_1 es la **pendiente** de la recta. Indica la variación media de la variable respuesta cuando X aumenta una unidad.
- El parámetro β_0 es el **término independiente** de la recta. Indica el valor medio de Y cuando $X = 0$.

Objetivo: estimar los parámetros β_0 y β_1 a partir de los datos (x_i, Y_i), $i = 1, \ldots, n$.

Datos con $\hat{\beta}_0 \approx 3$, $\hat{\beta}_1 \approx 0.5$ y $r \approx 0.8$



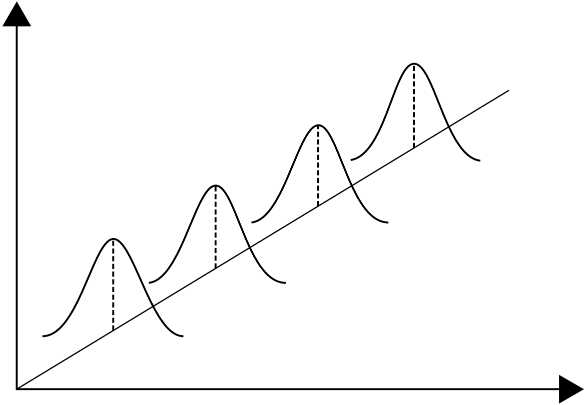
El modelo de regresión lineal simple

Para poder hacer inferencia (IC y contrastes) sobre los parámetros, suponemos que se verifica el siguiente modelo:
Para todas las observaciones $i = 1, \ldots, n$

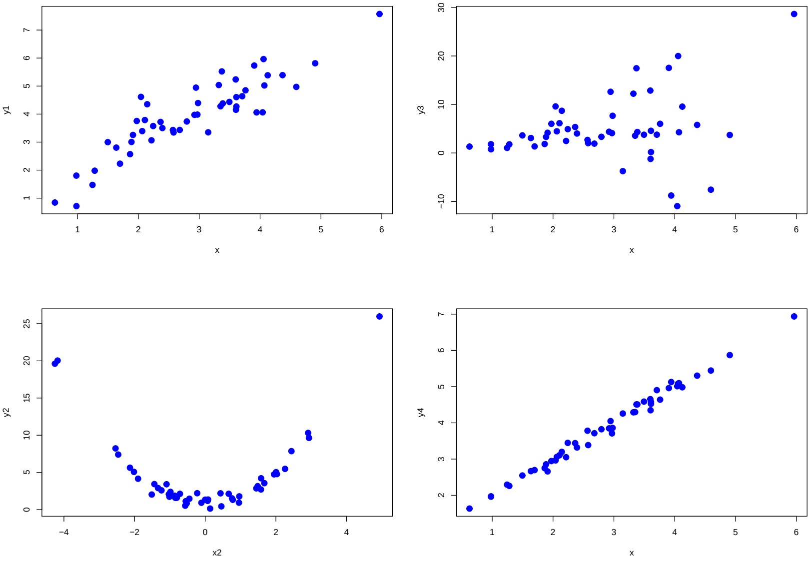
$$Y_i = \beta_0 + \beta_1 x_i + u_i,$$

donde:

- El valor medio de los errores u_i es cero.
- Todos los errores u_i tienen la misma varianza σ^2 (homocedasticidad).
- Las variables u_i tienen distribución normal.
- Las variables u_i son independientes.



¿En qué situaciones se verifica el modelo?



La recta de mínimos cuadrados

Si estimamos β_0 y β_1 mediante $\hat{\beta}_0$ y $\hat{\beta}_1$, la predicción de la variable respuesta Y_i en función de la regresora x_i es:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

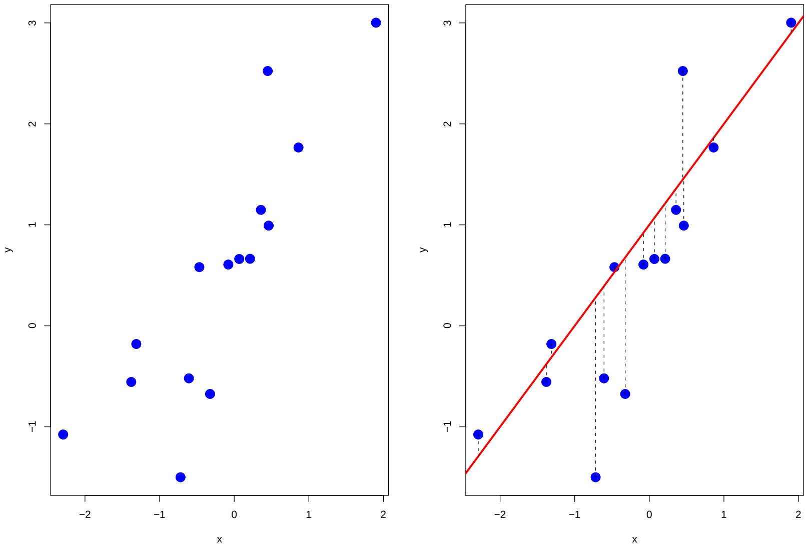
Unos buenos estimadores deben ser tales que los errores de predicción

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

sean pequeños.

La **recta de regresión de mínimos cuadrados** viene dada por los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ para los que se minimiza:

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



y
x
3
2
1
0
−1
−2
−1
0
1
2
y
x
3
2
1
0
−1
−2
−1
0
1
2

Estimadores de mínimos cuadrados

Pendiente:

$$\hat{\beta}_1 = \frac{\text{cov}_{xy}}{\text{v}_x} = r\frac{\sqrt{\text{v}_y}}{\sqrt{\text{v}_x}} = r\frac{S_y}{S_x}.$$

Término independiente:

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

Al igual que en los modelos de los temas anteriores:

- A las predicciones $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ se les llama **valores ajustados o pronosticados**.
- A los errores $e_i = Y_i - \hat{Y}_i$ se les llama **residuos**.

Ejemplo: temperatura y vibración de las alas

Estimadores de los parámetros:

$$\hat{\beta}_1 = r_{xy}\frac{S_y}{S_x} = 0.84\frac{1.73}{6.72} = 0.2155$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x} = 16.633 - 0.2155 \times 79.973 = -0.615$$

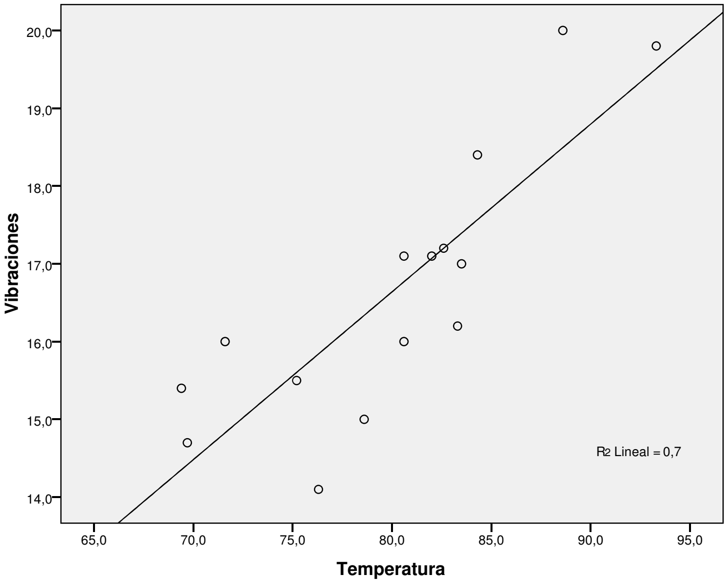
Recta de regresión:

$$y = -0.615 + 0.2155x$$

Predicción de Y_0 para $x_0 = 80$:

$$\hat{Y}_0 = -0.615 + 0.2155 \times 80 = 16.625$$

Diagrama de dispersión y recta estimada



Observaciones

- La recta de mínimos cuadrados pasa por el punto cuyas coordenadas son las medias: $(\bar{x}, \bar{Y})$.
- Si la variable regresora se incrementa en una desviación típica $\Delta x = S_x$, entonces la predicción de la variable respuesta se incrementa en r desviaciones típicas: $\Delta \hat{Y} = rS_y$
- Puede demostrarse que la suma de los residuos siempre vale cero.
- La recta para predecir Y en función de X no es la misma que la recta para predecir X en función de Y.

La varianza residual

La varianza residual es un estimador insesgado de σ^2:

$$S_R^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}.$$

Se pierden dos grados de libertad puesto que los residuos verifican dos restricciones:

- La media de los residuos es igual a cero.
- La covarianza entre los residuos y la variable regresora es también igual a cero.

Una simulación

Supongamos que $\sigma = 1$, $\beta_0 = 0$ y $\beta_1 = 1$.

Entonces el modelo es

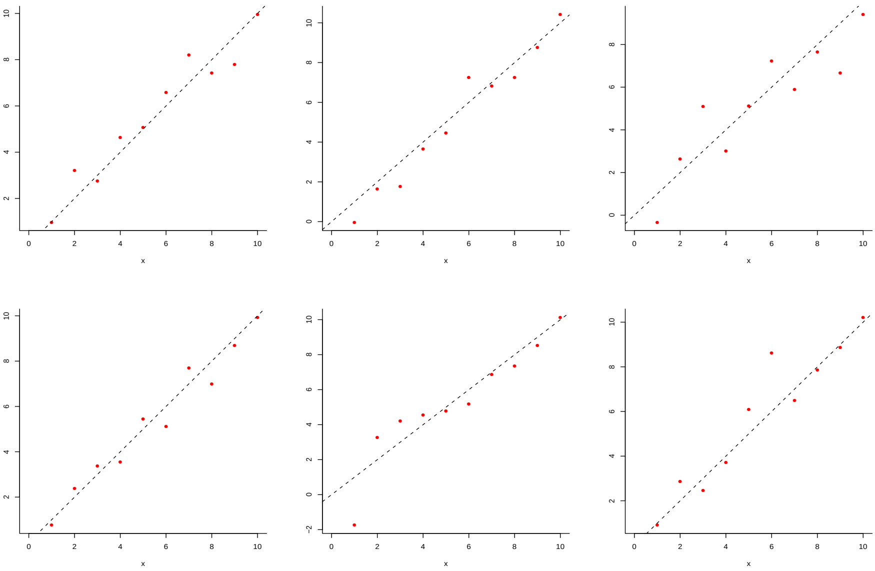
$$Y_i = x_i + u_i,$$

donde los errores u_i tienen distribución normal estándar y son independientes.

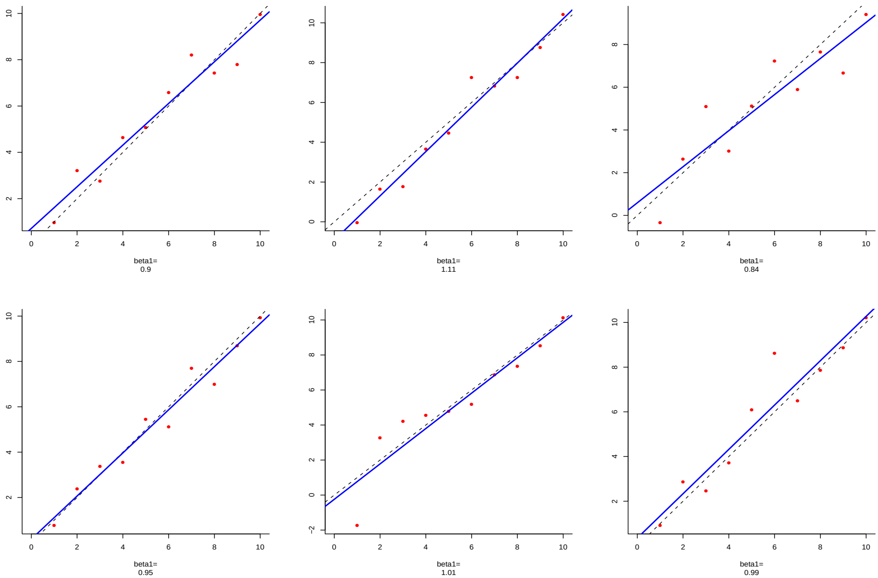
Fijamos $x_i = 1, 2, \ldots, 10$ ($n = 10$) y generamos las respuestas correspondientes de acuerdo con este modelo.

Posteriormente calculamos la recta de mínimos cuadrados y la representamos junto con la *verdadera recta* $y = x$.

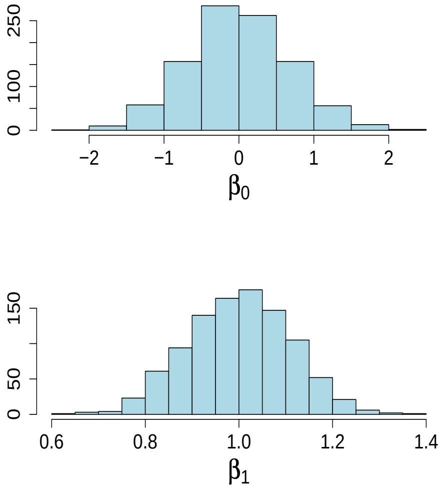
Repetimos 6 veces el experimento



Repetimos 6 veces el experimento



Repetimos 1000 veces el experimento



- Los estimadores son centrados y tienen distribución normal.
- Existen fórmulas del error típico de $\hat{\beta}_0$ y $\hat{\beta}_1$ que miden su variabilidad.
- Estas fórmulas son las que se utilizan para calcular IC y llevar a cabo contrastes en lo que sigue.

Error típico del estimador de la pendiente

$$\text{ERROR TÍPICO DE } \hat{\beta}_1 = \frac{S_R}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = S_R\sqrt{\frac{1}{n\mathrm{v}_x}}$$

- Al aumentar $n\mathrm{v}_x$, el error típico de la pendiente disminuye (es decir, la estimación de la pendiente es más precisa).
- Conviene diseñar el experimento de forma que los valores x_i tengan la mayor dispersión posible.

Error típico del estimador del término independiente

$$\textsc{error típico de } \hat{\beta}_0 = S_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n\mathrm{v}_x}}$$

- Si $\bar{x}^2$ es grande, se estima con menos precisión el término independiente.

Intervalos de confianza

Los intervalos de confianza de nivel $1-\alpha$ para los parámetros $\hat{\beta}_i$ $(i = 0, 1)$ tienen la estructura habitual:

$$\mathrm{IC}_{1-\alpha}(\beta_i) \equiv \left[\hat{\beta}_i \mp t_{n-2,\alpha/2} \times \text{ERROR TÍPICO DE } \hat{\beta}_i\right]$$

En comparación con los intervalos de confianza para la media:

- Los grados de libertad son $n-2$ en lugar de $n-1$.
- La fórmula del error típico es más complicada.

El intervalo de confianza para σ^2 también tiene la estructura que ya hemos visto en los modelos de los temas anteriores:

$$\mathrm{IC}_{1-\alpha}(\sigma^2) \equiv \left[\frac{(n-2)S_R^2}{\chi^2_{n-2;\alpha/2}}, \frac{(n-2)S_R^2}{\chi^2_{n-2;1-\alpha/2}}\right]$$

Ejemplo: temperatura y vibración de las alas

Para los datos del ejemplo se ha calculado $S_R^2 = 0.97$.

- Calcula los errores típicos de los estimadores de la pendiente y del término independiente.
- Calcula un intervalo de confianza de nivel 95% para β_1.
- Calcula un intervalo de confianza de nivel 95% para β_0.

Contrastes para los parámetros

Contraste bilateral:

- Hipótesis: $H_0 : \beta_i = 0$ frente a $H_1 : \beta_i \neq 0$
  Región crítica:

$$R = \left\{ \frac{|\hat{\beta}_i|}{\text{ERROR TÍPICO DE } \hat{\beta}_i} > t_{n-2,\alpha/2} \right\}.$$

Contrastes unilaterales:

- Hipótesis: $H_0 : \beta_i \leq 0$ frente a $H_1 : \beta_i > 0$
  Región crítica:

$$R = \left\{ \frac{\hat{\beta}_i}{\text{ERROR TÍPICO DE } \hat{\beta}_i} > t_{n-2,\alpha} \right\}.$$

- Hipótesis: $H_0 : \beta_i \geq 0$ frente a $H_1 : \beta_i < 0$
  Región crítica:

$$R = \left\{ \frac{\hat{\beta}_i}{\text{ERROR TÍPICO DE } \hat{\beta}_i} < -t_{n-2,\alpha} \right\}.$$

Ejemplo: temperatura y vibración de las alas

- ¿Aportan los datos evidencia para afirmar ($\alpha = 0.01$) que la temperatura tiene una influencia significativa sobre la frecuencia de vibración de las alas?

- ¿Podemos afirmar a nivel $\alpha = 0.01$ que al aumentar la temperatura aumenta la frecuencia media de vibración de las alas?

- Escribe la región crítica para contrastar $H_0 : \beta_1 = 1$ frente a $H_1 : \beta_1 \neq 1$.

Con SPSS: temperatura y vibraciones

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,836[a]	,700	,677	,9849

a. Variables predictoras: (Constante), Temperatura

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	29,383	1	29,383	30,290	,000[a]
	Residual	12,611	13	,970		
	Total	41,993	14			

a. Variables predictoras: (Constante), Temperatura

b. Variable dependiente: Vibraciones

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-,615	3,144		-,196	,848
	Temperatura	,216	,039	,836	5,504	,000

a. Variable dependiente: Vibraciones

Con SPSS: renta y fracaso escolar

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,742[a]	,550	,528	4,7566

a. Variables predictoras: (Constante), Renta

b. Variable dependiente: Fracaso

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	580,516	1	580,516	25,658	,000[a]
	Residual	475,133	21	22,625		
	Total	1055,649	22			

a. Variables predictoras: (Constante), Renta

b. Variable dependiente: Fracaso

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	38,494	3,645		10,562	,000
	Renta	-1,347	,266	-,742	-5,065	,000

a. Variable dependiente: Fracaso

Cuestiones

- Escribe la ecuación de la recta de mínimos cuadrados que describe el nivel de fracaso escolar como función de la renta.
- Calcula intervalos de confianza de nivel 95% para la pendiente y el término independiente de la recta de regresión.
- ¿Podemos afirmar, a nivel $\alpha = 0.05$ que niveles más altos de renta están asociados a niveles más bajos de fracaso escolar?
- ¿Cuánto vale el coeficiente de correlación entre el nivel de renta y el porcentaje de fracaso escolar?
- ¿Qué porcentaje de fracaso escolar se predice en una población cuya renta es $x_0 = 13000$ euros?
- ¿Cuál es el residuo correspondiente a Colmenar Viejo?

Análisis de la varianza en regresión simple

$$\begin{aligned}
Y_i &= \hat{Y}_i + e_i \\
Y_i - \bar{Y} &= (\hat{Y}_i - \bar{Y}) + e_i \\
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2 \\
\mathrm{SCT} &= \mathrm{SCE} + \mathrm{SCR}
\end{aligned}$$

SCT mide la variabilidad total (tiene $n-1$ gl)

SCE mide la variabilidad explicada por el modelo (tiene 1 gl)

SCR mide la variabilidad no explicada o residual (tiene $n-2$ gl)

Tabla ANOVA y contraste F

Fuente de variación	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada (SCE)	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	F
Residual (SCR)	$\sum_{i=1}^{n} e_i^2$	$n-2$	$S_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$	
Total (SCT)	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$		

El estadístico F es igual a SCE/S_R^2.

Si F es suficientemente grande (la variabilidad explicada es muy grande respecto a la no explicada), se debe rechazar $H_0 : \beta_1 = 0$.

Bajo $H_0 : \beta_1 = 0$, el estadístico F tiene distribución $F_{1,n-2}$. La región crítica de nivel α del contraste es:

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

Tabla ANOVA y contraste F

Para contrastar $H_0 : \beta_1 = 0$ a nivel α hemos considerado tres procedimientos:

- Calcular un IC de nivel de confianza $1 - \alpha$ para β_1 y rechazar H_0 si 0 no pertenece al intervalo.
- Dividir $|\hat{\beta}_1|$ por su error típico y rechazar H_0 si el valor obtenido es superior a $t_{n-2;\alpha/2}$.
- Calcular $F = \text{SCE}/S_R^2$ y rechazar H_0 si el valor obtenido es superior a $F_{1,n-2;\alpha}$.

Los tres métodos son equivalentes **en este modelo**.

Evaluación del ajuste

Para valorar el grado con el que la recta se ajusta a los datos se emplean varias medidas:

- El **coeficiente de correlación** r.

- El **coeficiente de determinación**:

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

  En el modelo de regresión simple $R^2 = r^2$, el coeficiente de determinación coincide con el coeficiente de correlación al cuadrado.

- El **error cuadrático medio**:

$$\text{ECM} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n}.$$

  Puede comprobarse que $\text{ECM} = V_y(1 - r^2)$.

Cuestiones

- Si SCT $= 8100$, SCE $= 6900$ y $\hat{\beta}_1 = -6.7$. Calcula el coeficiente de correlación entre la variable regresora y la variable respuesta.
- Para un conjunto de 20 datos se sabe que SCT $= 7200$, SCE $= 2900$ y $\hat{\beta}_1 = 3.1$. Calcula el coeficiente de correlación, el coeficiente de determinación y el error cuadrático medio.

Inferencia sobre la variable respuesta

Una de las razones para ajustar un modelo de regresión simple es obtener información sobre Y cuando x toma un valor x_0 conocido. Hay **dos problemas** relacionados con este objetivo:

- **Estimar el valor medio de** Y para los individuos de la población para los que $X = x_0$. Si μ_0 es este valor medio,

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

- **Predecir el valor individual que tomará la variable** Y para una nueva observación para la que se sabe que $X = x_0$. Si Y_0 es este valor,

$$Y_0 = \beta_0 + \beta_1 x_0 + u_0.$$

¿Qué problema es más difícil de los dos?
¿Qué estimador y qué predicción resultan razonables para μ_0 y Y_0?

Estimación y predicción puntual

En ambos casos, el estimador (o predicción) puntual es:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{Y} + \hat{\beta}_1 (x_0 - \bar{x}).$$

Sin embargo, el intervalo de confianza para μ_0 es diferente del intervalo de predicción para Y_0.

Intervalo de confianza para μ_0 de nivel $1 - \alpha$:

$$\left[\hat{Y}_0 \mp t_{n-2;\alpha/2} S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nV_x}} \right]$$

Intervalo de predicción para Y_0 de nivel $1 - \alpha$:

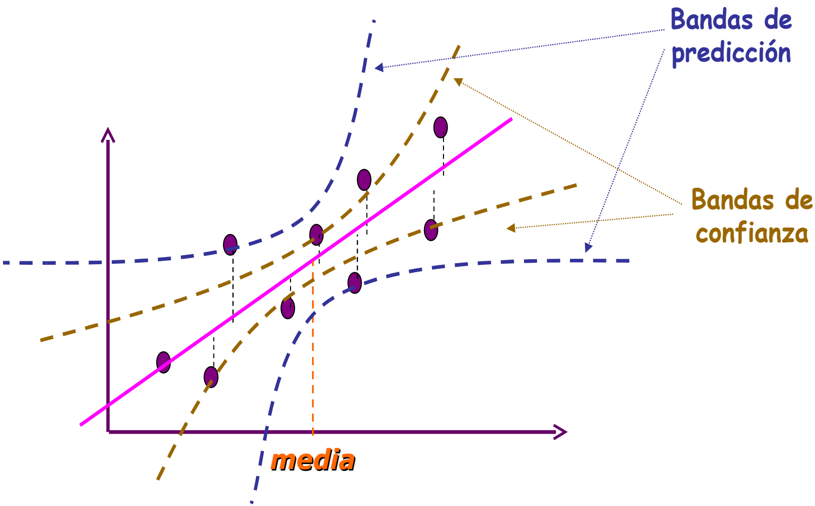
$$\left[\hat{Y}_0 \mp t_{n-2;\alpha/2} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{nV_x}} \right]$$

Ejemplo: temperatura y vibración de las alas

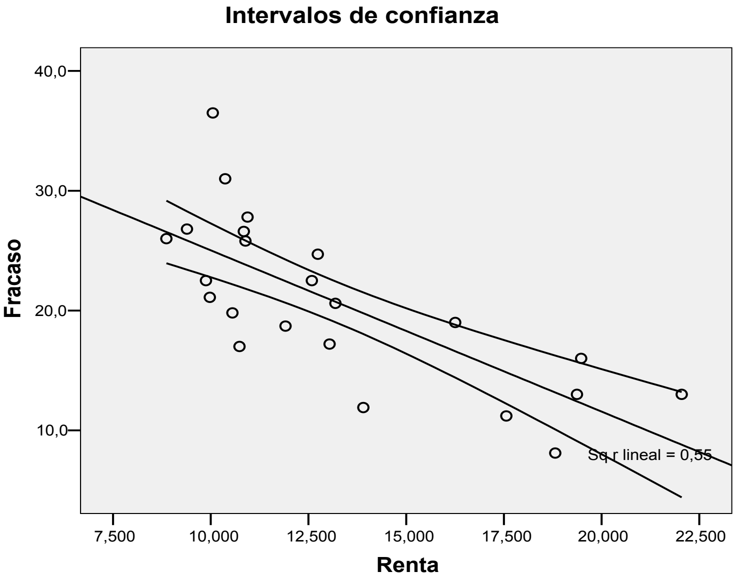
- Calcula un intervalo de confianza de nivel 95% para el número medio de vibraciones de las alas de los grillos cuando la temperatura es de 80 grados Farenheit.
- Calcula un intervalo de predicción de nivel 95% para el número de vibraciones de las alas de un grillo cuando la temperatura es de 80 grados Farenheit.
- En una población de la Comunidad de Madrid se sabe que la renta per cápita es 1000 euros inferior a la media de los datos disponibles. Calcula un intervalo de predicción de nivel 95% del porcentaje de fracaso escolar en esa población. Repite el ejercicio para una población cuya renta sea 1000 euros superior a la media.

	Medias	Cuasidesviaciones típicas
% Fracaso	20.73	6.92
Renta	13.19	3.81

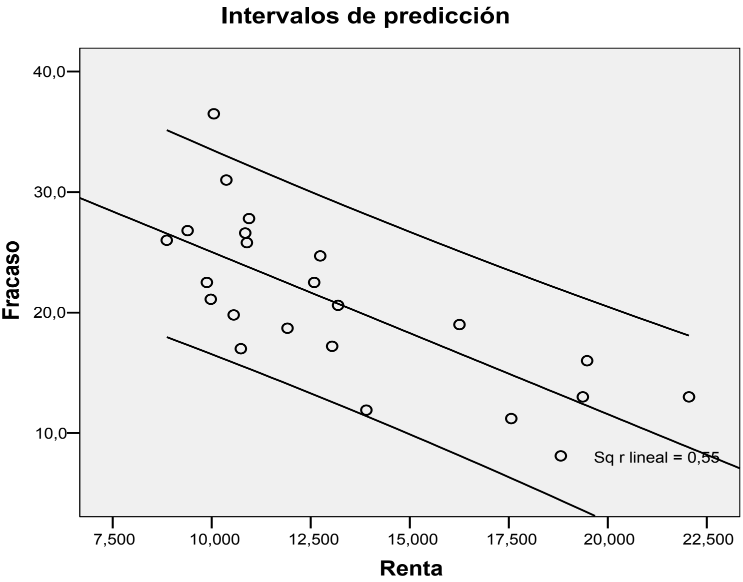
Intervalos de confianza y predicción



Intervalos de confianza para la media



Intervalos de predicción para valores individuales



Estimación de algunas relaciones no lineales

A veces, aunque la relación entre x e Y no sea lineal, el modelo de regresión simple puede aplicarse después de transformar adecuadamente las variables.

Modelos:

- Modelo de regresión exponencial
- Modelo de regresión logarítmica
- Modelo de regresión potencial

Modelo de regresión exponencial

La variable respuesta es aproximadamente una función exponencial de la variable regresora:

$$Y \approx ae^{bx}$$

Se linealiza tomando logaritmos:

$$\log Y \approx \log a + bx$$

Si ajustamos un modelo lineal a

$$(x_1, \log Y_1), \ldots, (x_n, \log Y_n)$$

obtenemos los estimadores $\widehat{\log a}$ y $\hat{b}$.

Invirtiendo los cambios obtenemos los estimadores $\hat{a}$ y $\hat{b}$.

Modelo de regresión logarítmica

La variable respuesta es aproximadamente una función lineal del logaritmo de la variable regresora:

$$Y \approx \beta_0 + \beta_1 \log x$$

Si ajustamos un modelo lineal a

$$(\log x_1, Y_1), \ldots, (\log x_n, Y_n)$$

obtenemos los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$.

Modelo de regresión potencial

La variable respuesta es proporcional a una potencia de la variable regresora:

$$Y \approx ax^b$$

Se linealiza tomando logaritmos:

$$\log Y \approx \log a + b \log x$$

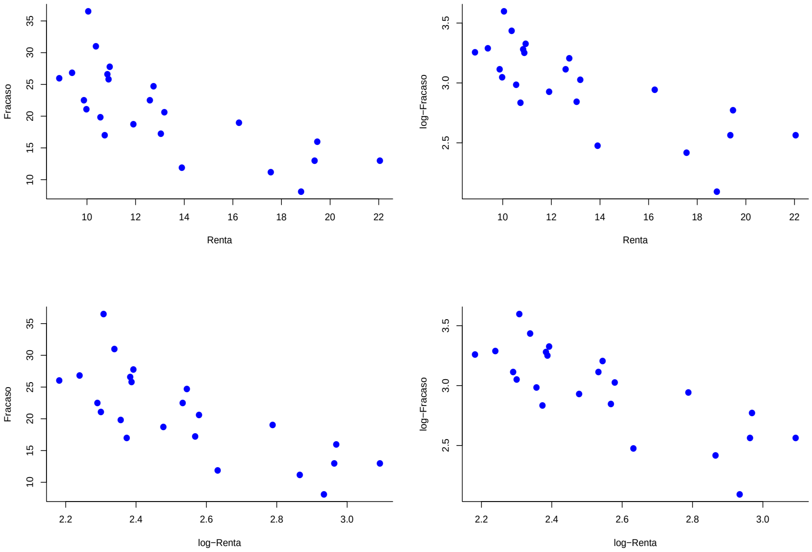
Si ajustamos un modelo lineal a

$$(\log x_1, \log Y_1), \ldots, (\log x_n, \log Y_n)$$

obtenemos los estimadores $\widehat{\log a}$ y $\hat{b}$.

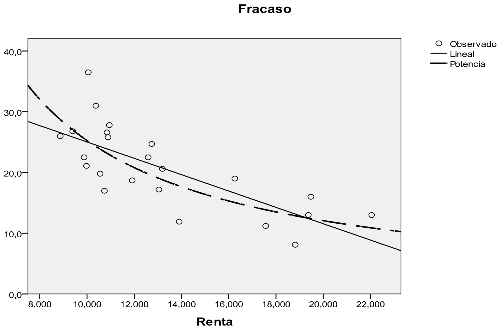
Invirtiendo los cambios obtenemos los estimadores $\hat{a}$ y $\hat{b}$.

Ejemplo: renta y fracaso escolar



Ejemplo: renta y fracaso escolar

Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,550	25,658	1	21	,000	38,494	-1,347
Logarítmica	,572	28,032	1	21	,000	70,584	-19,600
Potencia	,610	32,809	1	21	,000	293,923	-1,066
Exponencial	,594	30,691	1	21	,000	51,642	-,074



Diagnóstico del modelo: linealidad y homocedasticidad

El gráfico más útil para el diagnóstico del modelo es el de residuos frente a valores ajustados:

$$(\hat{Y}_1, e_1), \ldots, (\hat{Y}_n, e_n)$$

Se suelen utilizar los residuos estandarizados, que bajo las hipótesis del modelo tienen aproximadamente la distribución normal estándar.

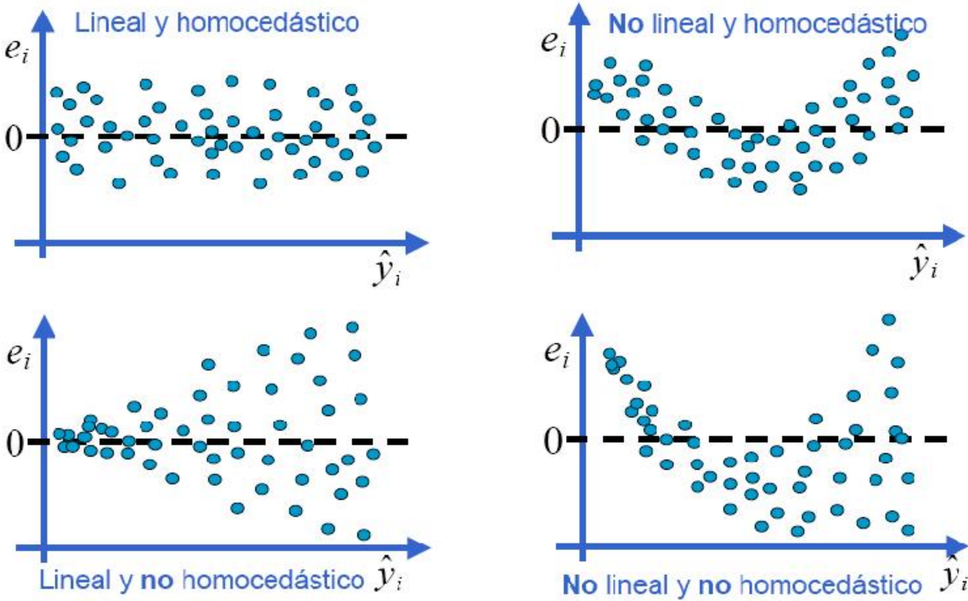
La hipótesis de normalidad se valora a partir de un gráfico de probabilidad de los residuos.

La homocedasticidad se puede confirmar si

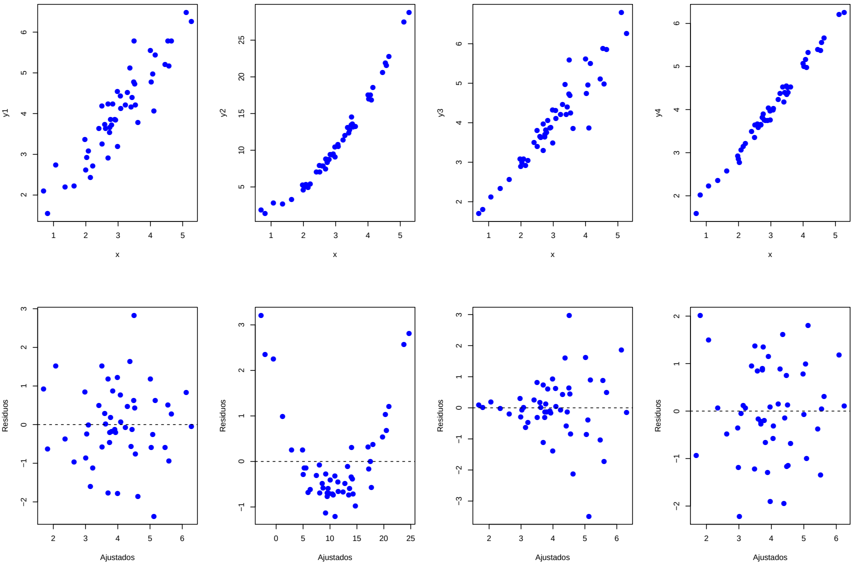
- No hay patrones sistemáticos en el gráfico.
- La variabilidad es aproximadamente constante a lo largo de todo el rango de valores ajustados.

Los residuos estandarizados que no están comprendidos entre los valores -3 y 3 pueden corresponder a datos atípicos potencialmente influyentes.

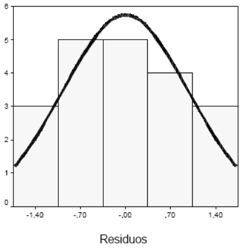
Residuos frente a valores ajustados

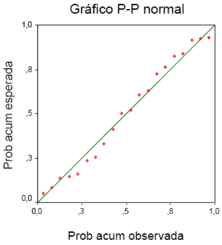


Residuos frente a valores ajustados



Diagnóstico del modelo: normalidad





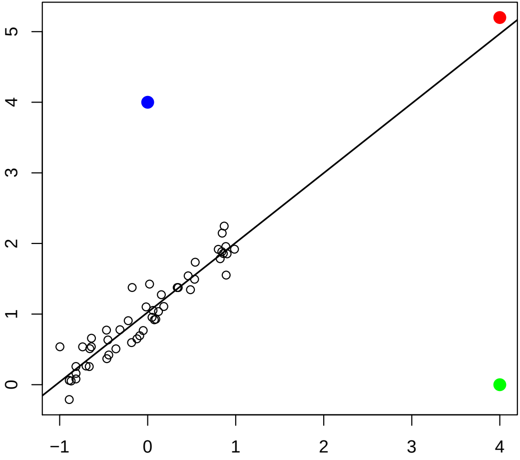
Prueba de Kolmogorov-Smirnov para una muestra

		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

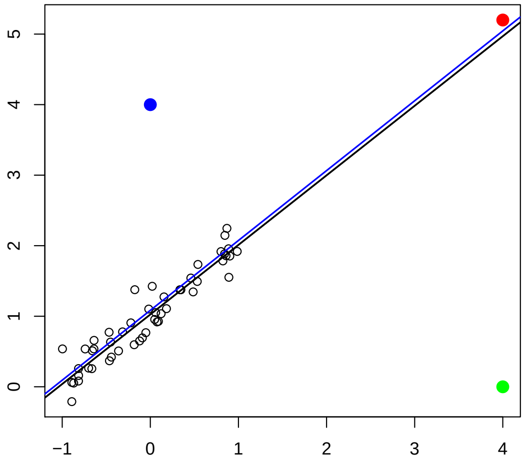
Precauciones al aplicar el modelo de regresión simple

- Existencia de datos atípicos
- Extrapolación
- Mezcla de poblaciones diferentes
- Datos temporales

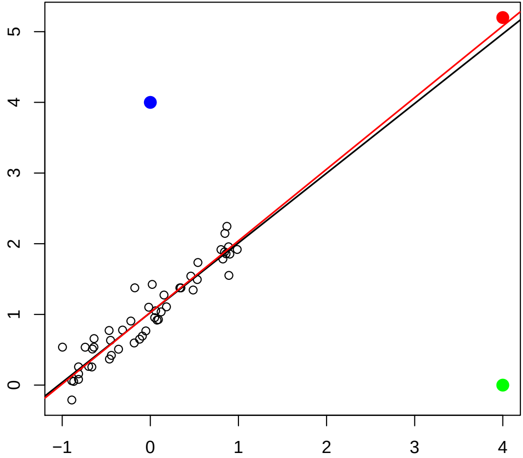
Datos atípicos



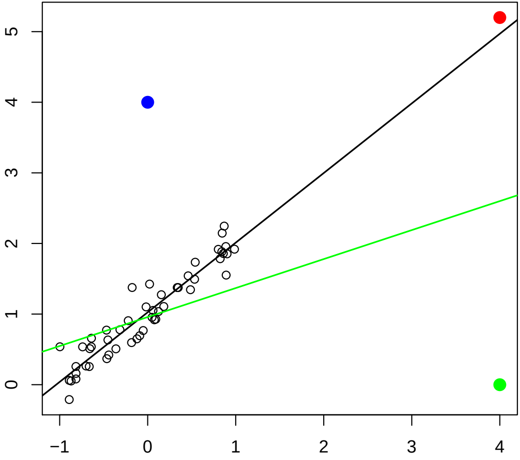
Datos atípicos



Datos atípicos

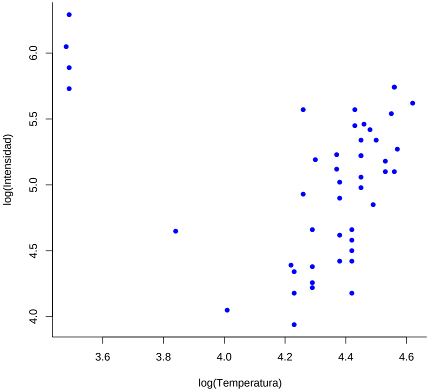


Datos atípicos

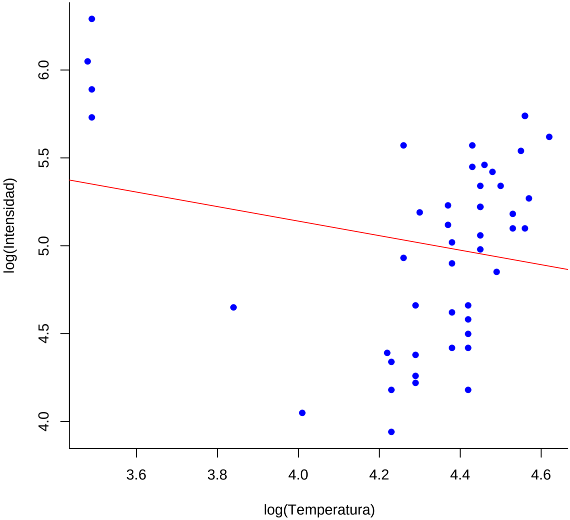


Ejemplo: Temperatura e intensidad de luz en estrellas

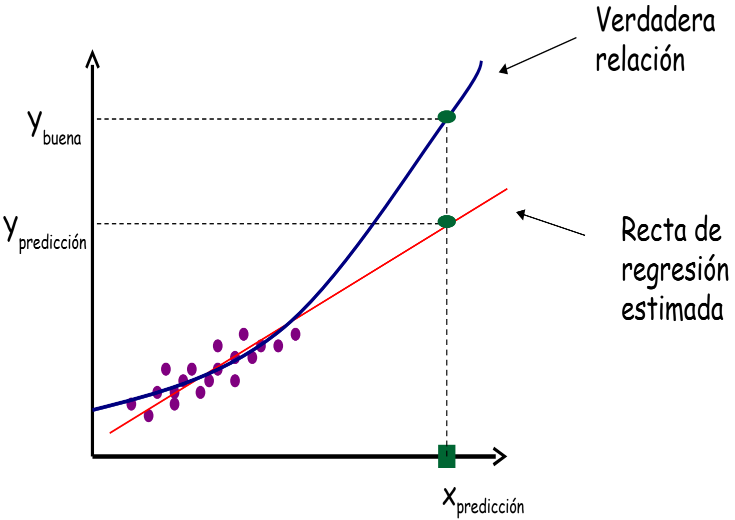
Para 47 estrellas se han registrado el log de la temperatura efectiva en la superficie (`Temp`) y el log de la intensidad de su luz (`Intens`).



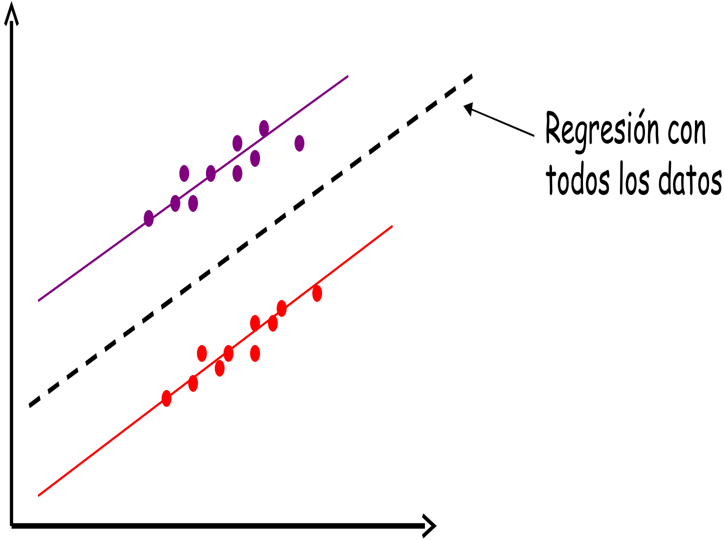
Ejemplo: Temperatura e intensidad de luz en estrellas



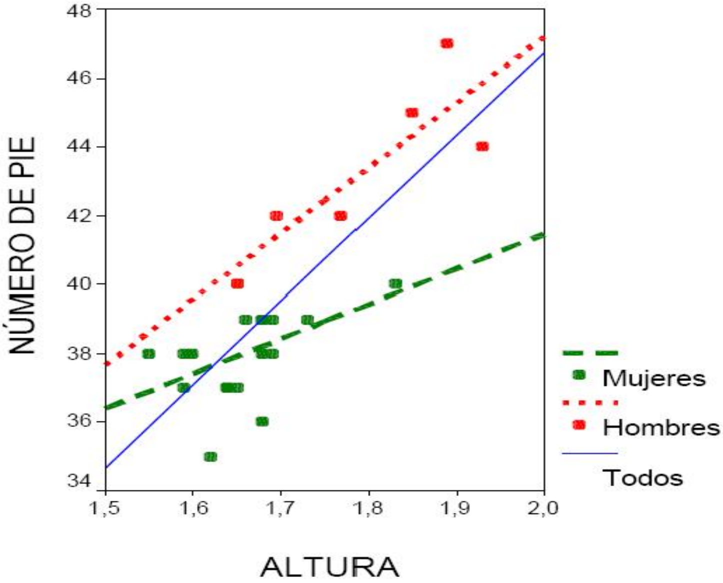
Extrapolación



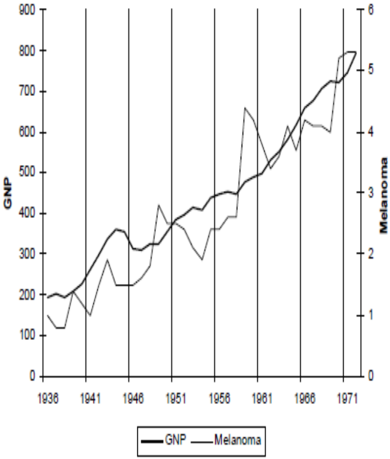
Mezcla de poblaciones

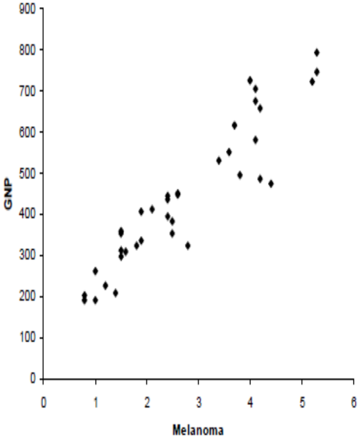


Ejemplo: número de pie y estatura



Datos temporales (correlación espúrea





PNB en EE.UU e incidencia del melanoma en la población masculina en Connecticut (1936-1972)