Regresión lineal simple y múltiple

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

September 25, 2019

Plan

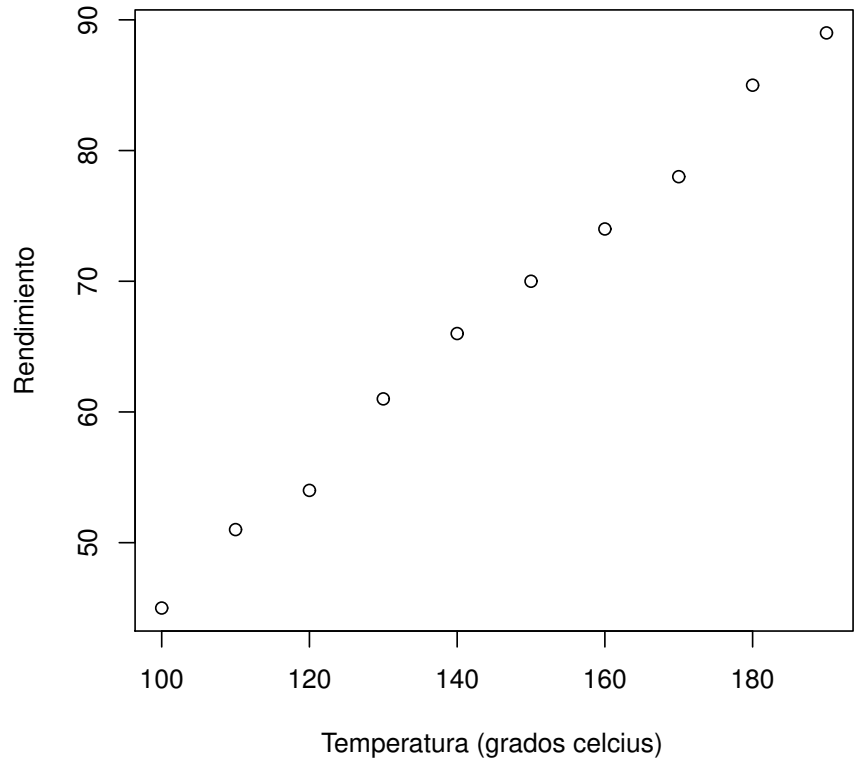
Plan

Regresión lineal simple. Primer Ejemplo

Objetivo: Establecer una relación entre una variable dependiente Y y una variable independiente x para poder hacer predicciones sobre Y cuando se conoce a x.

Ejemplo: Rendimiento de un producto químico en función de la temperatura.

Temp($^\circ C$)	Rend (%)
100	45
110	51
120	54
130	61
140	66
150	70
160	74
170	78
180	85
190	89



Se quiere expresar por medio de una ecuación la relación entre las variables x e y, mediante $y = f(x)$ con f a determinar. La gráfica sugiere una relación lineal.

El objetivo de la regresión lineal simple es de modelizar la variable aleatoria Y por una cierta función de X, $f(X)$ quién es la mejor en el sentido que minimiza el error cuadrático medio $\mathbb{E}\left((Y - f(X))^2\right)$.

Vimos en el curso anterior que esta función es la esperanza condicional de Y condicionada a X: $\mathbb{E}(Y|X)$. En el caso que las variables aleatorias son gausianas, el cálculo de la esperanza condicional da:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$

donde

$$\beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \qquad \beta_1 = \frac{Cov(X, Y)}{\text{Var}(X)}$$

La mejor función de X que permite modelizar Y es una función lineal de X, de donde el nombre de *regresión lineal*. Trataremos entonces de modelizar Y en función de X de manera lineal quién es la mejor modelización cuando las variables son gausianas. Trataremos de verificar que efectivamente la variable es gausiana, o en su defecto transformarla de manera que sea la más gausiana posible.

Si X e Y son independientes entonces la mejor modelización de Y en función de X es $\mathbb{E}(Y)$.

Regresión lineal simple. Primer ejemplo

Planteo del modelo lineal:

La obtención de una ecuación exacta $y = f(x)$ no siempre es posible e y puede depender de otros factores (*fenómenos aleatorios*). Se tendrá entonces un *error aleatorio* ϵ debido a variables y a factores no tenidos en cuenta, obteniendo de esta manera un modelo probabilístico para nuestro problema:

$$Y = f(x) + \epsilon$$

siendo ϵ el error aleatorio.

Volviendo a nuestro problema, nos proponemos hallar un modelo del tipo:

$$Y = \underbrace{\beta_0 + \beta_1 x}_{f(x)=x'\beta} + \epsilon$$

donde

- Y es la variable aleatoria dependiente, que se querrá predecir,
- x es la variable independiente, que se usa para predecir,
- β_0 y β_1 son parámetros desconocidos.
- ϵ es un error aleatorio.

Historia

TABLE XXII.

Father's Stature and Son's Stature.

Son's Stature.	Father's Stature. 58·5—59·5	59·5—60·5	60·5—61·5	61·5—62·5	62·5—63·5	63·5—64·5	64·5—65·5	65·5—66·5	66·5—67·5	67·5—68·5	68·5—69·5	69·5—70·5	70·5—71·5	71·5—72·5	72·5—73·5	73·5—74·5	74·5—75·5	Totals
59·5—60·5	—	—	—	—	·5	·5	1	—	—	—	—	—	—	—	—	—	—	2
60·5—61·5	—	—	—	—	·5	—	—	—	1	—	—	—	—	—	—	—	—	1·5
61·5—62·5	—	·25	·25	—	·5	1	·25	·25	·5	·5	—	—	—	—	—	—	—	3·5
62·5—63·5	—	·25	·25	2·25	2·25	2	4	5	2·75	1·25	—	·25	.25	—	—	—	—	20·5
63·5—64·5	1	—	1·5	3·75	3	4·25	8	9·25	3	1·25	1·5	·75	1·25	—	—	—	—	38·5
64·5—65·5	2	·1	·5	2	3·25	9·5	13·5	10·75	7·5	5·5	3·5	2·5	—	—	—	—	—	61·5
65·5—66·5	—	·5	1	2·25	5·25	9·5	10	16·75	17·5	16	5·25	2	2·5	1	—	—	—	89·5
66·5—67·5	—	1·5	2	4,75	3·5	13·75	19·75	26·5	25·75	19·5	12·5	13·75	3·25	·5	1	—	—	148·0
67·5—68·5	—	—	1·5	2	7·5	10	10·25	24·25	31·5	23·5	29·5	13·25	8·5	9·5	2·25	—	—	173·5
68·5—69·5	—	—	1	—	5·25	5	12·75	18·25	16	24	29	21·5	10	3·5	2·25	—	1	149·5
69·5—70·5	—	—	—	—	1	2·5	5·75	18·75	11·75	19·5	22·5	19·5	14·5	6·25	3·5	1·5	1	128·0
70·5—71·5	—	—	—	—	—	3·25	5	8·75	10·75	19	14·75	20·75	10·75	8	5	1	1	108·0
71·5—72·5	—	—	—	—	—	·25	3	1.25	7	7·75	10·75	11·25	10	8·5	2·75	·5	—	63·0
72·5—73·5	—	—	—	—	—	—	·75	·75	2·5	7·5	6·5	6	7·5	6·25	3·25	·5	·5	42·0
73·5—74·5	—	—	—	—	1	—	1·5	1·5	—	5·25	2·25	2·5	6·5	3·25	3·25	—	2	29·0
74·5—75·5	—	—	—	—	—	—	—	—	—	1	2	—	2·5	·75	1·75	·5	—	8·5
75·5—76·5	—	—	—	—	—	—	—	—	—	1·25	·25	—	·5	1	1	—	—	4·0
76·5—77·5	—	—	—	—	—	—	—	—	—	1·25	·25	1	—	—	1·5	—	—	4·0
77·5—78·5	—	—	—	—	—	—	—	—	—	—	1	1	—	·25	·75	—	—	3·0
78·5—79·5	—	—	—	—	—	—	—	—	—	—	—	—	—	·25	·25	—	—	·5
Totals	3	3·5	8	17	33·5	61·5	95·5	142	137·5	154	141·5	116	78	49	28·5	4	5·5	1078

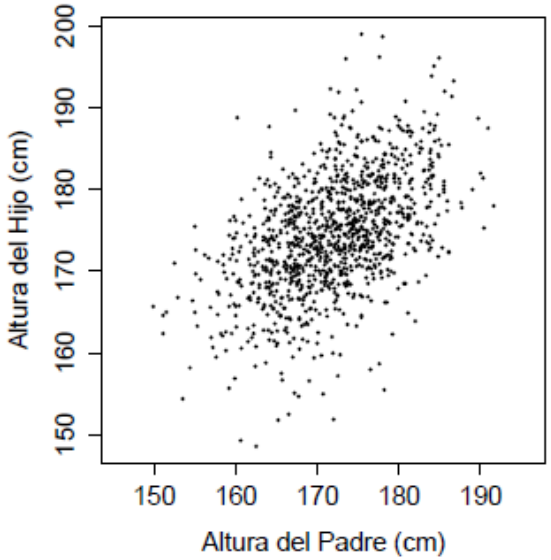
53—2

K. PEARSON AND A. LEE

415

Karl Pearson (1857-1936, matemático británico) observó la estatura de 1078 padres (x) e hijos (y).
Los promedios son $\overline{x} = 171,9$ cm e $\overline{y} = 174,5$ cm, los desvíos $s_x = 7$ cm y $s_y = 7.2$ cm, y $r = 0.5$

Historia



Observando que la recta de regresión se puede escribir como

$$y - \overline{y} = \widehat{\beta}_1(x - \overline{x})$$

se obtiene

$$y - \overline{y} = 0.51(x - \overline{x})$$

Si un padre tiene altura x, entonces

- Si $x > \overline{x}$ entonces $y > \overline{y}$ pero $y - \overline{y} < x - \overline{x}$.
- Si $x < \overline{x}$ entonces $y < \overline{y}$ pero $\overline{y} - y < \overline{x} - x$.

lo cual tiene la siguiente interpretación: los hijos cuyos padres tienen una estatura superior al valor medio, tienden a igualarse a éste, mientras que aquellos cuyos padres son muy bajos tienden a reducir su diferencia respecto a la estatura media, es decir, "regresan" al promedio.

Regresión lineal simple. Primer ejemplo

Planteo del modelo lineal:

Buscamos entonces la "mejor recta" según algún criterio de manera que pase lo más cerca posible de los puntos. En este contexto, el experto elije varios valores $x_1, \ldots, x_n$ de la variable X y observa los valores correspondientes $y_1, \ldots, y_n$ de la variable aleatoria Y.
Queremos hallar $\widehat{\beta_0}$ y $\widehat{\beta_1}$, estimadores de β_0 y β_1, que minimizan la suma de los errores cometidos al cuadrado:

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

e_i es la diferencia entre el valor y_i observado (donde "cae el punto") y el valor $\widehat{y_i}$ predecido por el modelo (donde "tendría que haber caído").

De esta manera, habiendo obtenido $\widehat{\beta_0}$ y $\widehat{\beta_1}$, para un valor x_0 de la variable independiente se podrá predecir por el modelo lineal el valor $\widehat{y_0}$ de la variable dependiente mediante

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

Regresión lineal simple. Método de los mínimos cuadrados

Método de los mínimos cuadrados

Una manera de minimizar el error $e_i = y_i - \widehat{y}_i$ consiste en minimizar la suma de los errores elevados al cuadrado, o la suma de los cuadrados residuales (SCR):

$$\text{SCR} = \sum_{i=1}^{n} e_i^2$$

Si el SCR es pequeño el ajuste es bueno, y si es grande el ajusto es malo.

En el caso de una recta vamos a querer hallar $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que minimicen

$$\sum_{i=1}^{n} (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2$$

Más adelante, veremos que si el gráfico de los puntos infieren que el modelo es cuadrático, vamos a querer hallar $\widehat{\beta}_0, \widehat{\beta}_1$ y $\widehat{\beta}_2$ que minimicen

$$\sum_{i=1}^{n} (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_2 x_i^2))^2$$

Regresión lineal simple. Método de los mínimos cuadrados

Método de los mínimos cuadrados

Derivamos $\sum_{i=1}^{n}(y_i-(\beta_1 x_i+\beta_0))^2$ respecto de β_1 y de β_0 e igualamos a 0:

$$\frac{\partial}{\partial\beta_1}\left(\sum_{i=1}^{n}(y_i-(\beta_1 x_i+\beta_0))^2\right)=-2\sum_{i=1}^{n}(y_i-(\beta_1 x_i+\beta_0))x_i=0$$

$$\frac{\partial}{\partial\beta_0}\left(\sum_{i=1}^{n}(y_i-(\beta_1 x_i+\beta_0))^2\right)=-2\sum_{i=1}^{n}(y_i-(\beta_1 x_i+\beta_0))=0$$

Despejamos β_0 de la primera ecuación y sustituyendo en la segunda obtenemos los estimadores MC (mínimos cuadrados) o LS (least squares):

$$\widehat{\beta_1}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{cov(x,y)}{s_x^2}=\underbrace{\frac{cov(x,y)}{s_y s_x}}_{r}\frac{s_y}{s_x}\qquad\widehat{\beta_0}=\overline{y}-\widehat{\beta_1}\overline{x}$$

donde $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$, $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $cov(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$

$s_x=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$, $s_y=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}$.

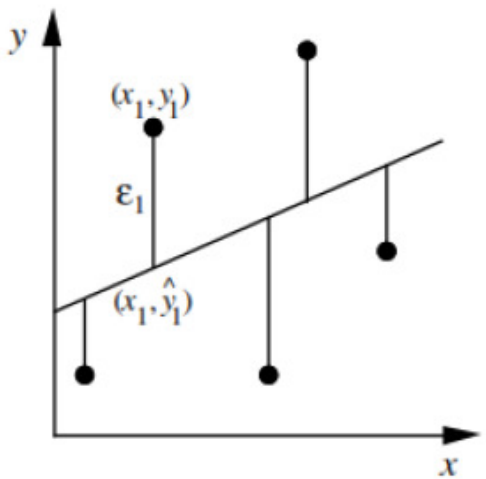
Es fácil ver que el punto encontrado es un mínimo.

Generalización

Ecuación fundamental:

$$\text{“observación”} = \text{“modelo”} + \text{“error aleatorio”}$$
$$Y = f(\mathbf{x}) + \epsilon$$

Los modelos de regresión utilizan la ecuación anterior suponiendo que el modelo es lineal.
En todo lo que sigue, consideramos una serie de datos $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:



"Mejor recta" $y = \beta_1 x + \beta_0$ de manera a minimizar $SCR = ||e||^2 = \sum_{i=1}^{n} e_i^2 = ||Y - X\beta||^2$

donde

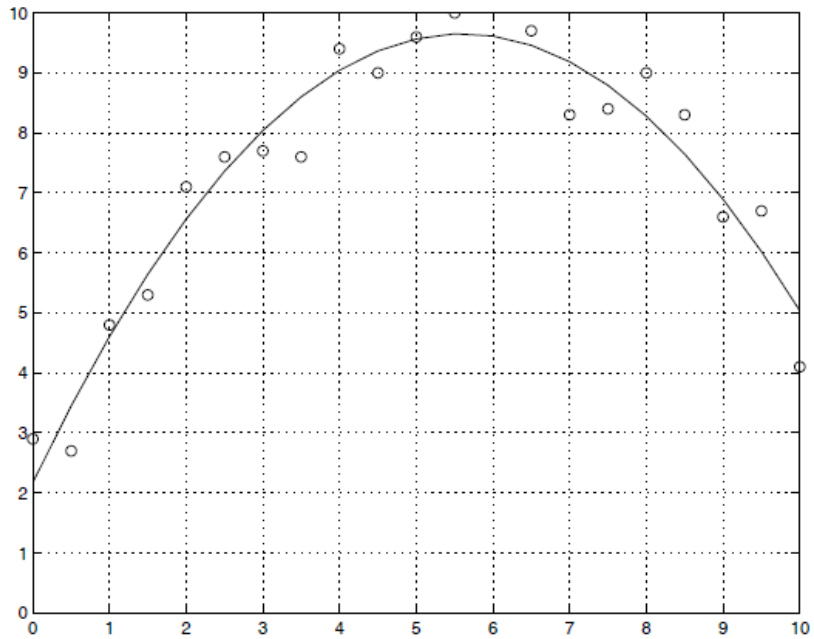
$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}_{e}$$

Generalización

Más generalmente podemos querer buscar el "mejor polinomio" de grado d

$$y = \beta_d x^d + \beta_{d-1} x^{d-1} + \cdots + \beta_1 x + \beta_0$$

que se ajusta a los datos.
Por ejemplo la parábola de mínimos cuadrados que ajusta un conjunto de puntos:



$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

(¡modelo lineal en los coeficientes!)

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_n & x_n^2 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix}}_{e}$$

Generalización

De la misma manera que para la regresión lineal simple, si $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ se quiere hallar un vector $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$ que minimice la función

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id})\right)^2$$

Hallamos entonces un hiperplano de regresión y podemos ver el problema como un problema de proyección ortogonal.

Observe que $\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id})\right)^2 = ||Y - X\beta||^2$ y por lo tanto el problema original se transforma en un problema de algebra lineal siendo:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}_{n \times (d+1)}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

Generalización - Visión geométrica

Si X es de rango completo, es decir $rg(X) = d + 1$ o $N(X) = \{0_{\mathbb{R}^{d+1}}\}$, entonces la solución por el método de los mínimos cuadrados es única, pues en este caso $X'X$ es invertible y por lo tanto

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

Si el rango de X es $r < d + 1$ entonces el sistema es indeterminado y la solución no es única y consideramos

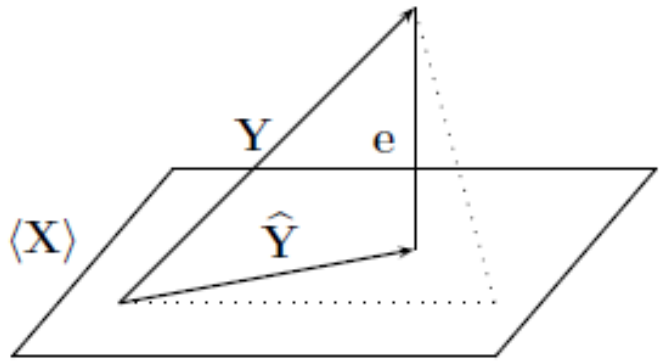
$$\widehat{\beta} = (X'X)^{-}X'Y$$

donde $(X'X)^{-}$ es una pseudo inversa de $X'X$ y verifica $(X'X)(X'X)^{-}(X'X) = (X'X)$.

Interpretación geométrica

$||e||^2 = e'e = ||Y - X\beta||^2$ es mínimo cuando

$$X\widehat{\beta} = P_{\langle X \rangle}(Y) = \widehat{Y}$$



Entonces

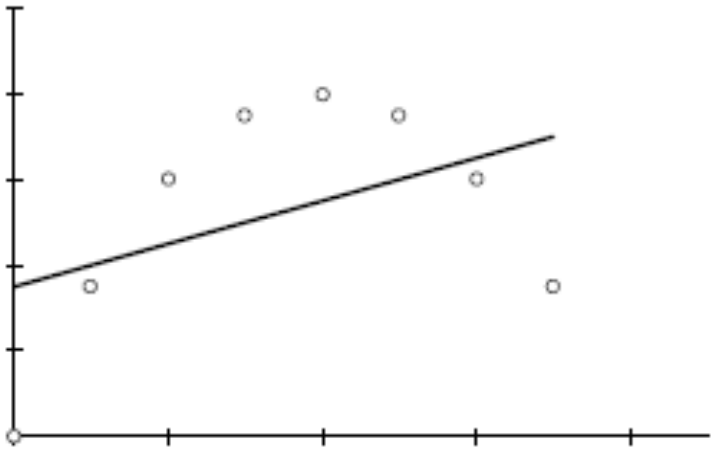
- $e = Y - \widehat{Y}$ es ortogonal a $\langle X \rangle$,
- $X'e = 0_{\mathbb{R}^{d+1}}$

Plan

Supuestos: condiciones de Gauss-Markov

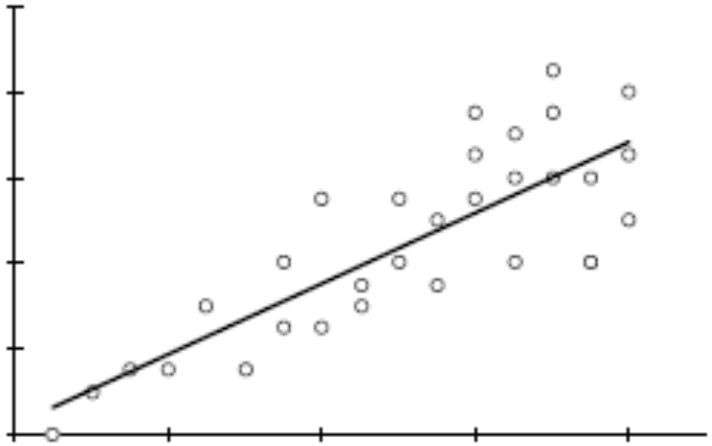
Hasta ahora el método de los mínimos cuadrados es análitico. Veamos donde interviene la estadística.
Suponemos que $x_1, \ldots, x_n$ son constantes. Supongamos que los errores e_i provienen de una variable aleatoria ϵ e impondremos que estos errores verifiquen las condiciones de Gauss-Markov:

(1) $\mathbb{E}(\epsilon_i) = 0$
$\Rightarrow \mathbb{E}(y_i) = \beta_1 x_i + \beta_0$
$\forall\, i = 1, \ldots, n$
No queremos que se dé esta situación:



Supuestos: condiciones de Gauss-Markov

(2) $Var(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \sigma^2$ (cte)
$\forall\, i = 1, \ldots, n$
(propiedad de homocedasticidad)
No queremos que se dé esta situación (heterocedasticidad):



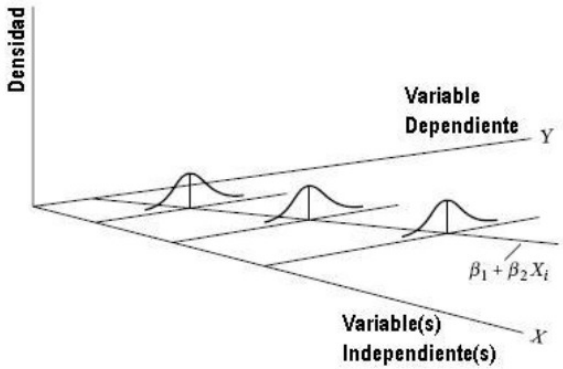


Figure: homocedasticidad

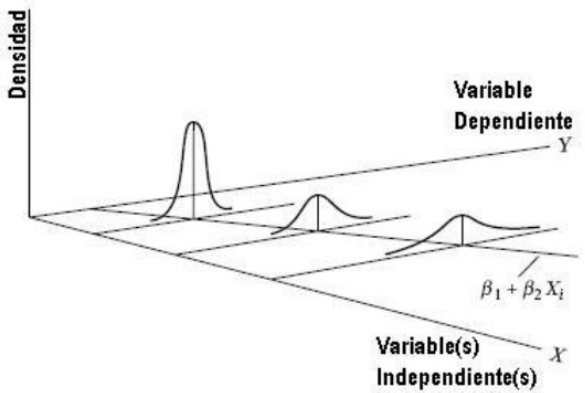


Figure: heterocedasticidad

Supuestos: condiciones de Gauss-Markov

(3) Los residuos deben ser incorrelados. Esto se puede hacer a partir del test de Durbin-Watson con el estadístico:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

que debe ser próximo a 2 si los residuos no son correlados. El estadístico no sigue ninguna ley en particular pero sus valores criticos están tabulados.

Resumen general

La expresión general del modelo lineal es:

$$Y = \underbrace{\mathbf{x}'\beta}_{f(\mathbf{x})} + \epsilon$$

y la estimación:

$$\widehat{\mathbf{y}} = \mathbf{x}'\widehat{\beta}$$

donde $\widehat{\beta}$ es la estimación del vector β obtenida por el método de los mínimos cuadrados.

Si suponemos las hipótesis de Gauss-Markov, el modelo lineal $Y = \mathbf{x}'\beta + \epsilon$ cumple que

$$\mathbb{E}(Y) = \mathbf{x}'\beta$$

Si además de suponer las condiciones de Gauss-Markov sobre los errores, se tiene que $\epsilon_i \sim N(0, \sigma^2)$ y que $\epsilon_1, \ldots, \epsilon_n$ son independientes, entonces decimos que el modelo es normal y se tiene que:

$$Y \sim N(\mathbf{x}'\beta, \sigma^2)$$

Esto último lo podemos verificar con un test de Shapiro Wilks.

Del modelo $Y = X\beta + \epsilon$, deducimos que matricialmente:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Entonces $(X'X)\beta = X'Y \Leftrightarrow \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n\overline{y} \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$

Por otro lado

$$(X'X)^{-1} = \frac{1}{ns_x^2} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

La recta de regresión en este caso es

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

siendo los estimadores:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$$

La recta de regresión se expresa también como

$$y - \overline{y} = \widehat{\beta}_1(x - \overline{x})$$

y por lo tanto para todo $i = 1, \ldots, n$ se tiene que $\widehat{y_i} - \overline{y} = \widehat{\beta}_1(x_i - \overline{x})$ y por lo tanto $\sum(\widehat{y_i} - \overline{y})^2 = \widehat{\beta}_1^2 s_x$.

- Se prueba que $\widehat{\beta_0}$ y $\widehat{\beta_1}$ son estimadores insesgados de β_0 y β_1 y de varianza mínima.
- La estimación para x_i es:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

- la diferencia entre $\widehat{y_i}$ e y_i es el residuo $\widehat{e_i} = \widehat{y_i} - y_i$
- La varianza residual σ^2 se estima por

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

Muchos modelos no lineales se pueden transformar en modelos lineales mediante transformaciones sencillas:

- $Y = \alpha X^{\beta}$
- $Y = \alpha e^{\beta X}$
- ...

Se recomiendo representar la nube de puntos (x_i, y_i) para darse cuenta.

Regresión lineal simple. Primer ejemplo

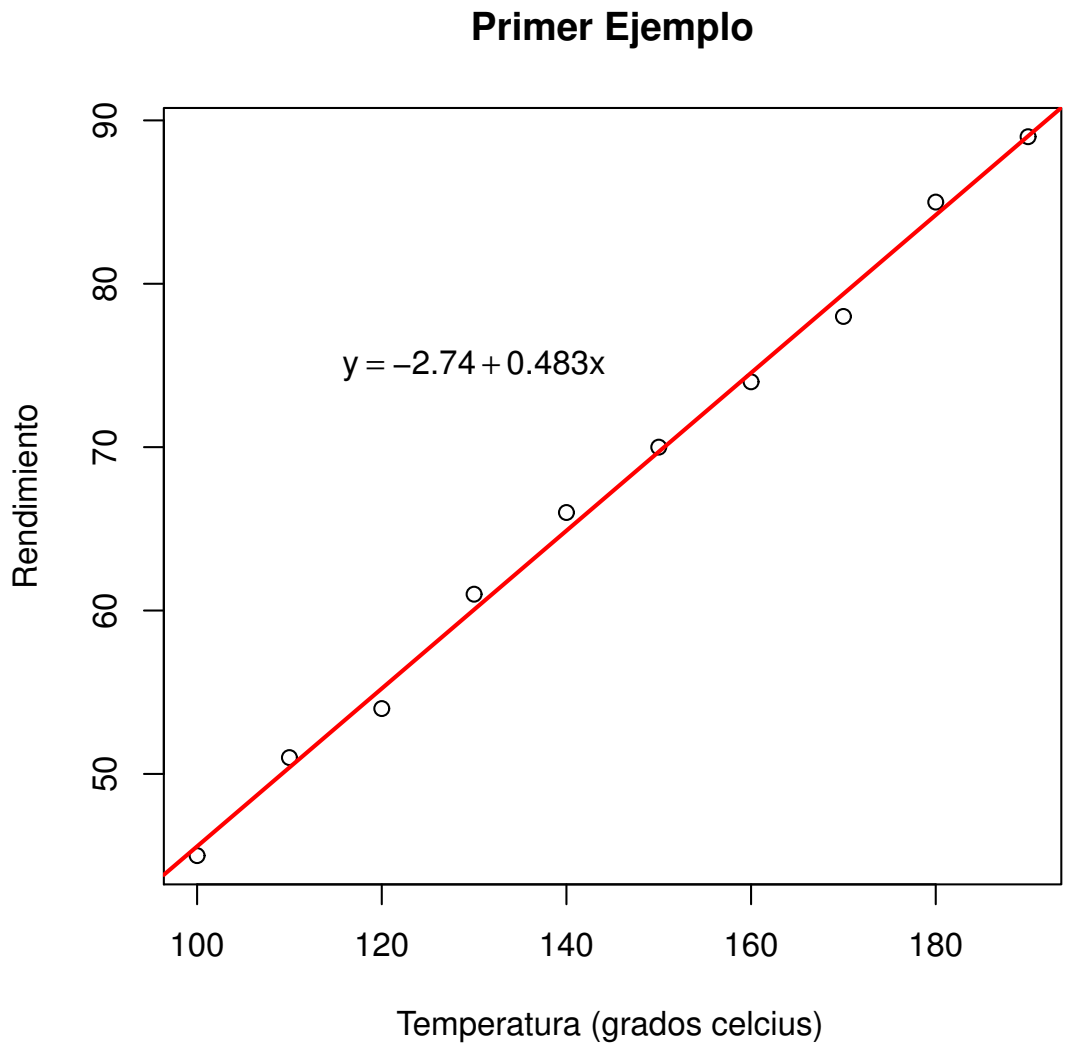
Volvemos a nuestro problema inicial:

```
>X=cbind(seq(100,190,10),c(45,51,54,61,66,70,74,78,85,89))
> X=as.data.frame(X)
> colnames(X)=c("Temp","Rend")
> plot(X,xlab="Temperatura (grados celcius)",ylab="Rendimiento",
main=paste("Primer Ejemplo"))
> a=lm(Rend~Temp,data=X)
> summary(a)
> abline(a,col="red",lwd=2)
Call:
lm(formula = Rend ~ Temp, data = X)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3758 -0.5591  0.1242  0.7470  1.1152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.73939    1.54650  -1.771    0.114
Temp         0.48303    0.01046  46.169 5.35e-11 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.9503 on 8 degrees of freedom
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9958
F-statistic:  2132 on 1 and 8 DF,  p-value: 5.353e-11
```

Regresión lineal simple. Primer ejemplo

La ecuación de la recta es

$$\widehat{y} = -2,74 + 0.48x$$



La función lm

Vamos a tratar de entender un poco más esta función y de ver las distintas posibilidades de hacer regresión lineal.

La linealidad es sobre **los coeficientes** del modelo, es decir, el modelo es lineal en los parámetros $\beta_0, \beta_1, \ldots, \beta_d$ que se quiere hallar:

1. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ es lineal
2. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}^2 + \beta_4 x_{i2} x_{i4} + e_i$ es lineal
3. $y_i = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \cos(x_{i2}) + \beta_3 x_{i3}^2 + \beta_4 x_{i2} x_{i4} + e_i$ es lineal.
4. $y_i = \beta_0 + \beta_1 \sin(\beta_2 x_{i1}) + \beta_2 x_{i2}^{\beta_3} + e_i$ NO es lineal.

Con el R, las funciones que se usan son:

```
>lm(y~x1+x2) #para el modelo y=ax1+bx2+c
>lm(y~I(x1+x2)) #para el modelo y=a(x1+x2)+c
>lm(y~poly(x,2)) #para el modelo y=ax^2+bx+c
>lm(y~x-1) #para el modelo y=ax
```

Ejemplo

Se quiere modelar la relación que existe entre el salario Y (en millones de dolares) y la cantidad de años de experiencia x de profesionales y obtener un intervalo de confianza al 95% para Y cuando $x = 10$.

Nuestra base de datos consiste de 143 observaciones:

```
>profsalary <- read.table("profsalary.txt",header=TRUE)
>attach(profsalary)
>plot(Experience,Salary,xlab="Years of Experience", main=paste("Salary data"))
```


```
> head(profsalary,10)
   Case Salary Experience
1     1     71         26
2     2     69         19
3     3     73         22
4     4     69         17
5     5     65         13
6     6     75         25
7     7     66         35
8     8     66         16
9     9     67         16
10   10     69         16
```

Ejemplo

Claramente esta relación no es lineal y no sería adecuada el modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 x + e$$

siendo Y el salario y x la cantidad de años de experiencia. Claramente el ploteo sugiere un modelo de regresión polinomial cuadrático

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

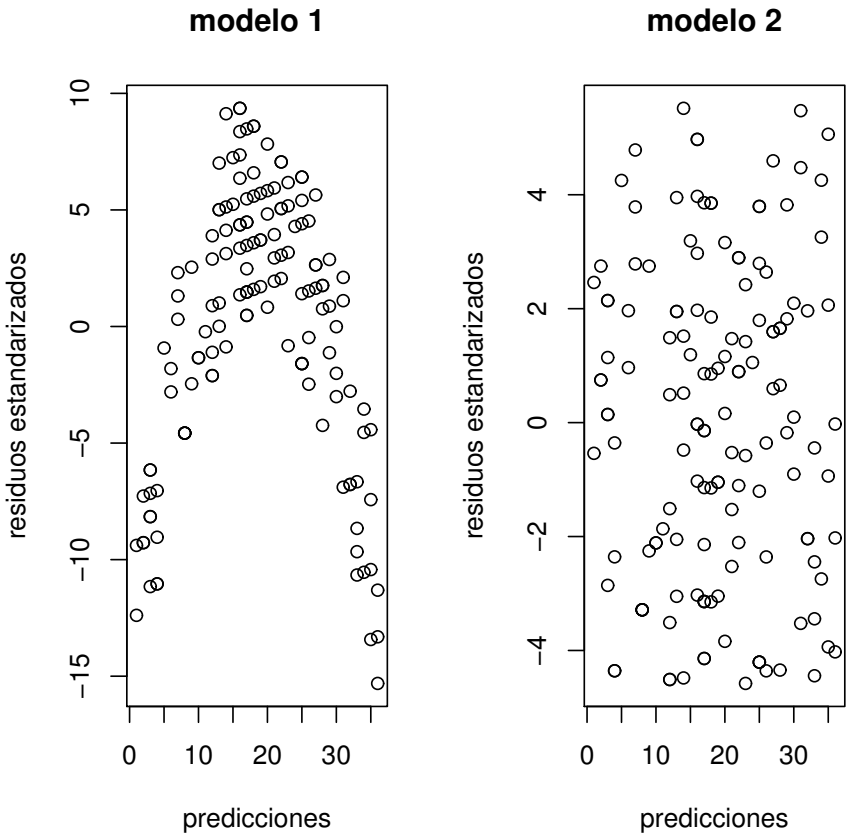
```
>m1 <- lm(Salary~Experience)
>abline(m1,col="red",lwd=2)
```


```
>m2 <- lm(Salary~Experience +
I(Experience^2))
```

Ejemplo

Acá vamos a graficar los errores (estandarizados) cometidos por cada modelo.

```
>par(mfrow=c(1,2))
>plot(Experience,m1$res,xlab="predicciones",
ylab="residuos estandarizados",main=paste("modelo 1"))
>plot(Experience,m2$res,xlab="predicciones",
ylab="residuos estandarizados",main=paste("modelo 2"))
```



El segundo modelo parecería más adecuado: no hay patrón en cuanto a los errores cometidos.

Ejemplo

```
> summary(m2)

Call:
lm(formula = Salary ~ Experience + I(Experience^2))

Residuals:
    Min      1Q  Median      3Q     Max
-4.5786 -2.3573  0.0957  2.0171  5.5176

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.720498   0.828724   41.90   <2e-16 ***
Experience       2.872275   0.095697   30.01   <2e-16 ***
I(Experience^2) -0.053316   0.002477  -21.53   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.817 on 140 degrees of freedom
Multiple R-squared:  0.9247,    Adjusted R-squared:  0.9236
F-statistic: 859.3 on 2 and 140 DF,  p-value: < 2.2e-16

>
```

Bajo la hipótesis de normalidad de los residuos, los estimadores $\widehat{\beta_0}$ y $\widehat{\beta_1}$ de β_0 y β_1 tienen distribución

$$\widehat{\beta_1} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_X^2}\right) \quad \text{y} \quad \widehat{\beta_0} \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \overline{x}^2}{S_X^2}\right)$$

donde $S_X^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$ y estimamos la varianza por $\widehat{\sigma^2} = \frac{SCR}{n-2}$.

Se prueba que:

- $\frac{n-2}{\sigma^2} SCR \sim \chi^2_{n-2}$
- $\frac{\widehat{\beta_0} - \beta_0}{s.e(\widehat{\beta_0})} \sim t_{n-2}$ donde $s.e(\widehat{\beta_0})^2 = var(\widehat{\beta_0}) = \widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_x}\right) = \frac{SCR}{n-2}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_x}\right)$
- $\frac{\widehat{\beta_1} - \beta_1}{s.e(\widehat{\beta_1})} \sim t_{n-2}$ donde $s.e(\widehat{\beta_1})^2 = var(\widehat{\beta_1}) = \widehat{\sigma}^2 \frac{1}{S_x} = \frac{SCR}{n-2}\frac{1}{S_x}$

Esto permite construir intervalos de confianza y de testear la nulidad de los parámetros.

Tabla de significancia modelo

En la regresión lineal simple, se quiere testear si hay relación de linealidad entre Y y X. El test es:

$$\begin{cases} H_0 : \text{No hay relación lineal} \\ H_1 : \text{Hay relación lineal} \end{cases}$$

Source	grados libertad	Sum. Squares	Mean Square	F
Modelo	1	$SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	MSR=SSR/1	MSR/MSE
Error	$n-2$	$SCR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	MSE= SCR/$(n-2)$	
Total	$n-1$	$SST = S_y = \sum_{i=1}^{n} (y_i - \overline{y})^2$		

El estadístico $F = MSR/MSE$ con el que se testea la hipótesis nula $\beta_1 = 0$ contra la hipótesis $\beta_1 \neq 0$ tiene distribución F con 1 y $n-2$ grados de libertad.

Un valor de MSE pequeño indica que el model ajusta bien ($\widehat{y}_i \approx y_i$), en cambio un valor grande de MSE indica que el modelo no sería razonable.

Se rechaza H_0 si $F > F_\alpha(1, n-2)$.

Regresión lineal simple. Inferencias sobre los paramétros

Supongamos el modelo $Y = \beta_0 + \beta_1 X + \epsilon$

Prueba de hipotesis sobre la pendiente
Con hipótesis de normalidad sobre los residuos se testea:

$$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases}$$

cuyo estadśitico es $T_1 = \frac{\widehat{\beta}_1 - b_1}{s.e(\widehat{\beta}_1)}$.

Región critica: $\left| \frac{\widehat{\beta}_1 - b_1}{s.e(\widehat{\beta}_1)} \right| > t_{n-2}(\alpha/2)$.

Observación: En el caso $b_1 = 0$, con un $p-$valor pequeño podemos inferir que existe una relación entre Y y X. O sea, un resultado significativo que rechace H_0 puede implicar que el modelo lineal sea adecuado, pero podría ser que no lo sea igual (no confundir significación de la regresión con causalidad). Por otro lado, es equivalente al test F, pues

$$T_1 = \frac{\widehat{\beta}_1}{s.e(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{\sqrt{\frac{SCR}{(n-2)S_x}}} = \frac{\widehat{\beta}_1 \sqrt{S_x}}{\sqrt{\frac{SCR}{(n-2)}}} = \sqrt{\frac{SSR}{MSE}} = \sqrt{F}$$

Intervalo de confianza al $100(1-\alpha)\%$ **para** β_1**:**

$$\left[\widehat{\beta}_1 - t_{n-2}(\alpha/2)s.e(\widehat{\beta}_1), \widehat{\beta}_1 + t_{n-2}(\alpha/2)s.e(\widehat{\beta}_1)\right]$$

Regresión lineal simple. Inferencias sobre los paramétros

Prueba de hipotesis sobre el intercepto
Con hipótesis de normalidad:

$$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases}$$

Región critica: $\left| \frac{\widehat{\beta}_0 - b_0}{s.e(\widehat{\beta}_0)} \right| > t_{n-2}(\alpha/2)$.

Intervalo de confianza al $100(1-\alpha)\%$ **para** β_0**:**

$$\left[\widehat{\beta}_0 - t_{n-2}(\alpha/2)s.e(\widehat{\beta}_0), \widehat{\beta}_0 + t_{n-2}(\alpha/2)s.e(\widehat{\beta}_0)\right]$$

Intervalo de confianza al $100(1-\alpha)\%$ **para** σ^2**:**
Se prueba que un estimador para σ^2 es $\widehat{\sigma}^2 = \frac{SCR}{n-d-1}$. Como $SCR/\sigma^2 \sim \chi^2_{n-2}$, se tiene que:

$$\left[\frac{SCR}{\chi^2_{n-2}(\alpha/2)}, \frac{SCR}{\chi^2_{n-2}(1-\alpha/2)}\right]$$

Regresión lineal simple. Descomposición variación

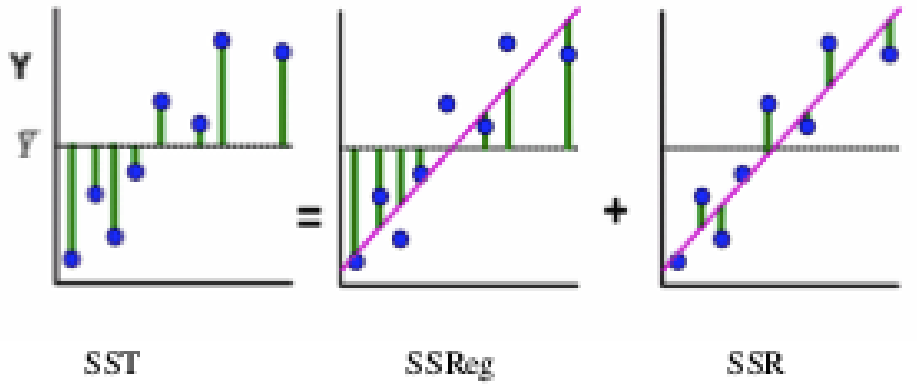
Con nuestras notaciones, si $\widehat{y}_i$ es la predicción de x_i por el modelo, se verifica lo que llamamos la *descomposición de la variación:*

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{Variación total VT o SST}} = \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{Variación no explicada VNE o SCR}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{\text{Variación explicada VE o SSReg}}$$

En efecto:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\widehat{y}_i+\widehat{y}_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2 + 2\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y})}_{0}$$

porque $\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y}) = \sum_{i=1}^{n}(y_i-\widehat{y}_i)\widehat{y}_i - \overline{y}\sum_{i=1}^{n}(y_i-\widehat{y}_i) = \sum_{i=1}^{n} e_i\widehat{y}_i - \overline{y}\sum_{i=1}^{n} e_i = 0-0=0$



Regresión lineal simple. Coeficiente de determinación R^2

La proporción de variabilidad explicada por el modelo es el *coeficiente de determinación* :

$$R^2 = \frac{VE}{VT} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{VT - VNE}{VT} = 1 - \frac{SCR}{S_y}$$

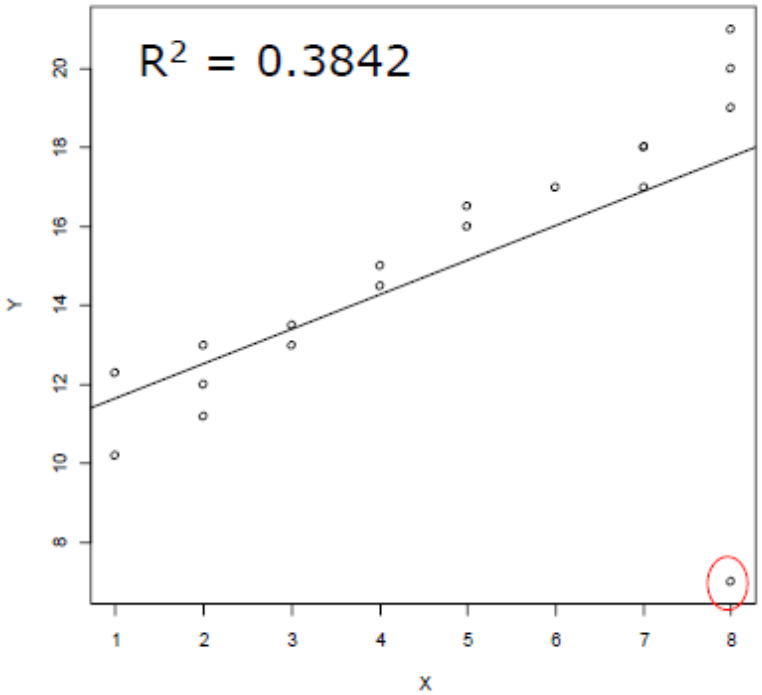
El coeficiente de determinación R^2 es una medida de la bondad del ajuste, **suponiendo que el modelo es lineal**. *En el caso de la regresión lineal simple* coincide con r^2.

- Observar que $0 \leq R^2 \leq 1$: si el valor de R^2 es cercano a 1 entonces gran parte de la variabilidad es explicada por el modelo, mientras que si está cerca de 0, una parte importante de la variabilidad no está explicada por el modelo (es probable que el modelo no sea adecuado).
- Cuidado que el R^2 no es una medida de adecuación del modelo. Es una medida de cuan significativo es el modelo una vez que establecimos que responde a un modelo lineal. Para ver si el modelo se ajusta a un modelo lineal, se usa el test Lack of Fit (LOF) cuando tenemos réplicas.
- Puede ocurrir también que la presencia de algún outlier implique que R^2 es bajo y hacernos pensar que el modelo no es bueno cuando en realidad sí lo es.
- Para corregir el peligro de sobreajuste se define el coeficiente de determinación ajustado como

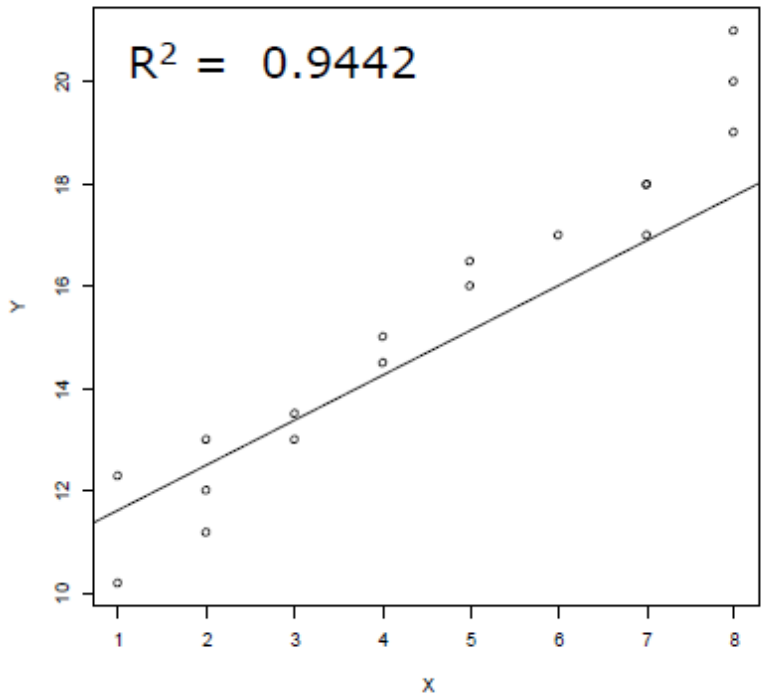
$$\overline{R}^2 = 1 - \frac{SCR/(n-2)}{S_y/(n-1)}$$

Si R^2 y $\overline{R}^2$ son muy distintos es que el modelo fue sobreajustado e inducirnos a mirar de más cerca las variables y/o cambiar la cantidad de términos.

Regresión lineal simple. Coeficiente de determinación R^2



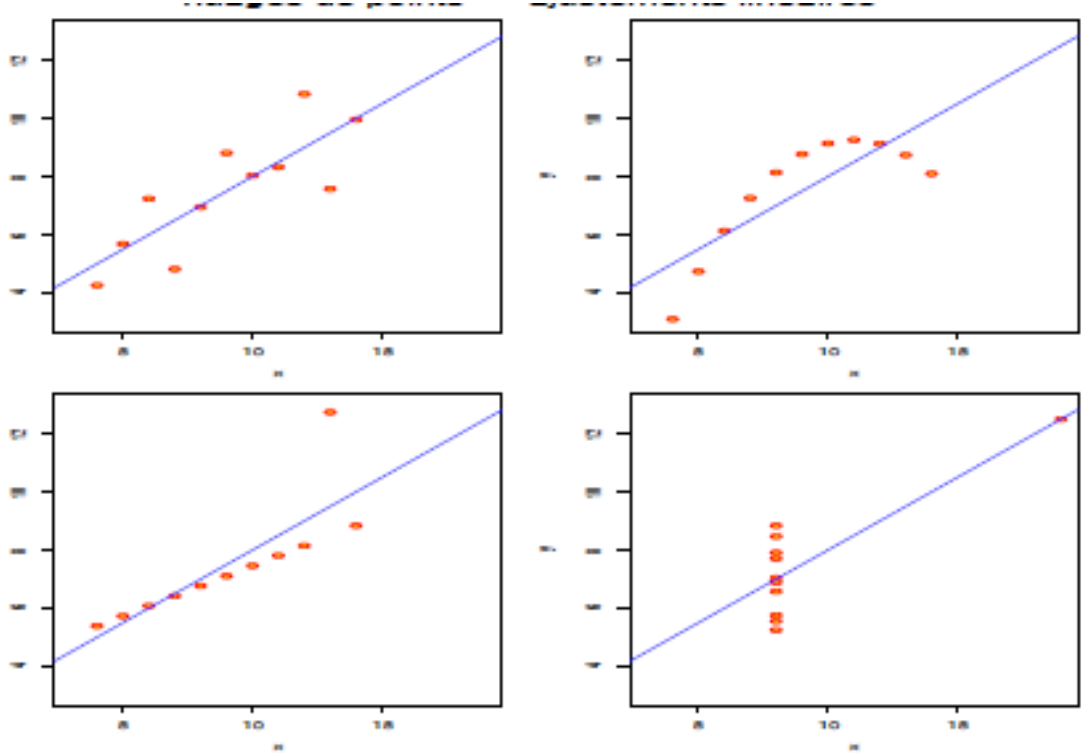
Comportamiento del R^2 con y sin un dato «outlier» en la variable Y.



Regresión lineal simple. Coeficiente de determinación R^2

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	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
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

$$\bar{x} = 9\,;\ \bar{y} = 7.50,$$
$$S_x^2 = 10\,;\ S_y^2 = 3.75$$
$$r = 0.816.$$



Plan

Intervalo de confianza para la recta de regresión

Intervalo de confianza para respuesta media:
Se trata de un intervalo de confianza para $\mathbb{E}(Y|X = x_0)$, la respuesta media al valor x_0.

Dado un valor determinado x_0 de la variable independiente, como el error tiene una distribución $\mathcal{N}(0, \sigma^2)$, la variable $Y = \beta_0 + \beta_1 x_0 + \epsilon$ tiene distribución $\mathcal{N}(\mu, \sigma^2)$ donde la media μ de Y es $\mu = \beta_0 + \beta_1 x_0$

Para un x_0 dado, consideramos el pronóstico $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$. El valor $\widehat{y}_0$ es un estimador de $\mu = \mathbb{E}(Y|x_0))$.
Un intervalo de confianza al nivel $1 - \alpha$ para la respuesta media $\mu = \beta_0 + \beta_1 x_0 = \mathbb{E}(Y|x_0)$ es

$$\left[\widehat{y}_0 - t_{\alpha/2,n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x} \right)}, \widehat{y}_0 + t_{\alpha/2,n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x} \right)} \right]$$

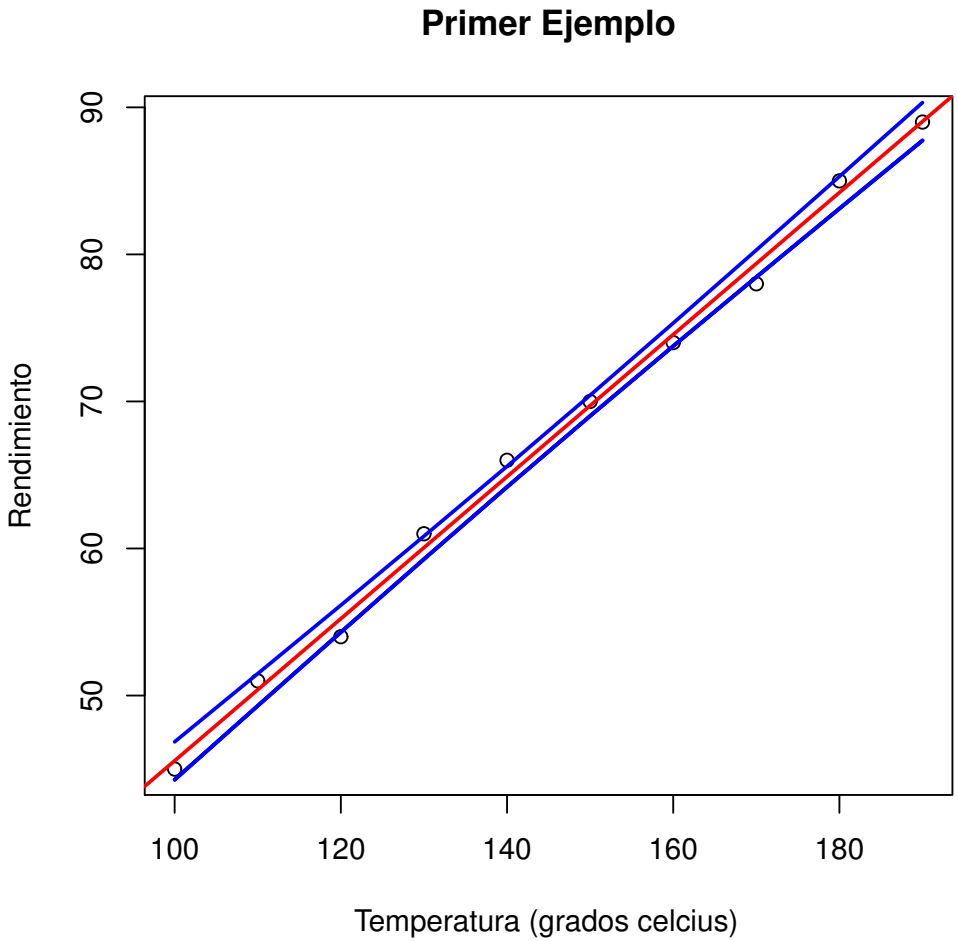
El ancho del intervalo depende de x_0, es mínimo cuando $x_0 = \overline{x}$ y crece cuando $|x_0 - \overline{x}|$ crece.

Intervalo de confianza para la recta de regresión

:

Intervalo de confianza para respuesta media $IC_{1-\alpha}\left(\mathbb{E}(Y|X=x_0)\right)$:

x_0	100	110	120	130	140	150	160	170	180	190
y	45	51	54	61	66	70	74	78	85	89
$\widehat{y}_0$	45.56	50.39	55.22	60.05	64.88	69.72	74.55	79.38	84.21	89.04
límites	$\pm$1.30	$\pm$1.10	$\pm$0.93	$\pm$0.79	$\pm$0.71	$\pm$0.71	$\pm$0.79	$\pm$0.93	$\pm$1.10	$\pm$1.30



Intervalo de confianza para la predicción

El intervalo definido anterioremente es adecuado para el valor esperado de la respuesta, pero ahora queremos un intervalo de predicción para una respuesta individual concreta.

Intervalo de confianza para la predicción: $IC_{1-\alpha}(y_0)$
Sea y_0 el verdadero valor (desconocido por lo tanto) de Y cuando la variable independiente es x_0. $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ es estimador puntual de un nuevo valor de la respuesta $Y_0 = Y|x_0$. Si consideramos un intervalo de confianza para esta futura observación Y_0, el intervalo de confianza para la respuesta media en $x = x_0$ no es apropiado ya que es un intervalo sobre la media de Y_0 (un parámetro), y no sobre futuras observaciones de la distribución. La variable $Y_0 - \widehat{y}_0 \sim \mathcal{N}\left(0, \text{Var}(Y_0 - \widehat{y}_0)\right)$ donde

$$\text{Var}(Y_0 - \widehat{y}_0) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_x} \right)$$

pues Y_0, una futura observación, es independiente de $\widehat{y}_0$.

Un intervalo de confianza al nivel $1 - \alpha$ para y_0 es

$$\left[\widehat{y}_0 - t_{\alpha/2, n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x} \right)}, \widehat{y}_0 + t_{\alpha/2, n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x} \right)} \right]$$

Consideraciones finales

- Los radios de los dos intervalos crecen cuando x_0 se aleja de $\overline{x}$.
- Los intervalos de predicción para una nueva observación son más amplios que los intervalos de confianza para los parámetros desconocidos. El tamaño del intervalo de confianza para un parámetro depende de la incertidumbre de la estimación que hacemos a partir de una muestra. Mientras que el tamaño del intervalo de predicción para una nueva observación tiene dos fuentes de incertidumbre: una debida a la estimación de los parámetros desconocidos y la otra es propia de la aleatoriedad que suponemos (es una variable aleatoria!).

Plan

Diagnóstico del modelo

Los métodos de estimación son muy sensibles a los outliers. Una vez detectado un outlier, no hay solución obvia, todo depende del contexto (se puede sacarlo, ver si se debe a un error de medición o no hacer nada).

- residuo $e_i = \widehat{y}_i - \widehat{y}_i$
- Se puede probar que $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$ donde h_{ii} es la entrada de la diagonal de la matriz H llamada *hat matriz* (la que "pone el gorro" en la Y), la matriz de proyección sobre $\langle X \rangle$, es decir $\widehat{Y} = HY$.

  Se puede probar que

$$\widehat{Y}_i = \sum_{i=1}^{n} h_{ij} Y_j$$

  donde

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{y} \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  Los terminos h_{ii} dan una medida del impacto de y_i en la estimación de $\widehat{y}_i$. Este impacto está directamente relacionado con el alejamiento de x_i con $\overline{x}$.

- Muchas veces se recomienda trabajar con los *residuos estandarizados (internos)*

$$r_i = \frac{e_i}{\sqrt{\widehat{\sigma^2}(1 - h_{ii})}} \quad \forall\, i = 1, \ldots, n$$

Diagnóstico del Modelo

- Otra alternativa consiste en calcular los *residuos studentizados externamente* o *R-Student*:

$$t_i = \frac{e_i}{\sqrt{s^2_{(i)}(1-h_{ii})}} \sim t_{n-3}$$

donde $s^2_{(i)}$ es una estimación de σ^2 que se calcula con las $n-1$ observaciones sin tener en cuenta $\mathbf{x}_i$ y se demuestra que:

$$s^2_{(i)} = \frac{(n-2)\widehat{\sigma} - \frac{e_i^2}{(1-h_{ii})}}{n-3}$$

En la mayoría de las situaciones, t_i no difiere mucho de r_i. Sin embargo, si la i-esima obervación es influyente, $s^2_{(i)}$ puede diferir significativamente de $\widehat{\sigma^2}$ y el estadístico t_i ser más sensible en este punto.

Así la varianza de un error e_i depende de la posición del punto $\mathbf{x_i}$ al punto central $\bar{\mathbf{x}}$: puntos cercanos a $\bar{\mathbf{x}}$ tienen mayor varianza (pobre ajuste por mínimos cuadrados) que los puntos lejanos.

En la práctica, un diagnóstico a ojo es más rápido. Se considera en general, residuo atípico o *outlier* si $|t_i| > 2$.

- Si no hay explicación aparente frente a un outlier se debe hacer el análisis con y sin él, a la espera de nuevos datos, o alguna explicación adicional
- Si no se elimina.

Nivel de un punto

Muchas veces se puede observar que algún dato o un subconjunto de datos ejerce una influencia muy grande sobre el modelo de regresión, es decir que el modelo depende más de estos datos atipicos que de la mayoría de los puntos. A estos puntos se le llaman *puntos influyentes*. Vamos a querer localizarlos y ver cual es el impacto en el modelo.

El *nivel de un punto* o *leverage* h_{ii} de $\mathbf{x_i}$ es una medida de distancia del punto al centroide $\bar{\mathbf{x}}$. Como traza$(H) = d + 1$ entonces el tamaño medio de cada h_{ii} es $\frac{d+1}{n}$. Cuando un punto $\mathbf{x_i}$ tenga $h_{ii} > 2\frac{d+1}{n}$ diremos que es un punto con nivel alto, o con *leverage* alto. Estos puntos se deben observar para su posterior estudio dado que son potencialmente influyentes. También se pueden observar aquellos puntos que tienen leverage mayor a los otros.

- Se puede probar que

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \quad \text{(ésto sale de } \widehat{Y} = HY\text{)}$$

  Entonces si h_{ii} está cerca de 1 se tiene que los h_{ij}'s son pequeños e y_i contribuye más a la determinación de $\widehat{y}_i$ que los otros y's. Los puntos que tienen *leverage* elevado son puntos que muy posiblemente van a ser determinante en cuanto a la regresión. En general, si la observación i tiene *leverage* cercano a 1, entonces la estimación estará cercana a y_i y por lo tanto $e_i = y_i - \widehat{y}_i$ será chico.

  Los terminos h_{ii} se obtienen a partir de la función hatvalues()

- Tener *leverage* alto puede ser bueno o malo.
- Combinación peligrosa: observación con *leverage* alto y un residuo studentizado alto.

Gráficos asociados a los errores

Recordamos que las hipótesis sobre los residuos son:

1. $\mathbb{E}(\epsilon) = 0_{\mathbb{R}^n}$.
2. $\text{Var}(\epsilon) = \sigma^2 I_n$.
3. $\epsilon_i \sim \mathcal{N}(0, \sigma^2) \ \forall\, i = 1, \ldots, n$.

Podemos hacer:

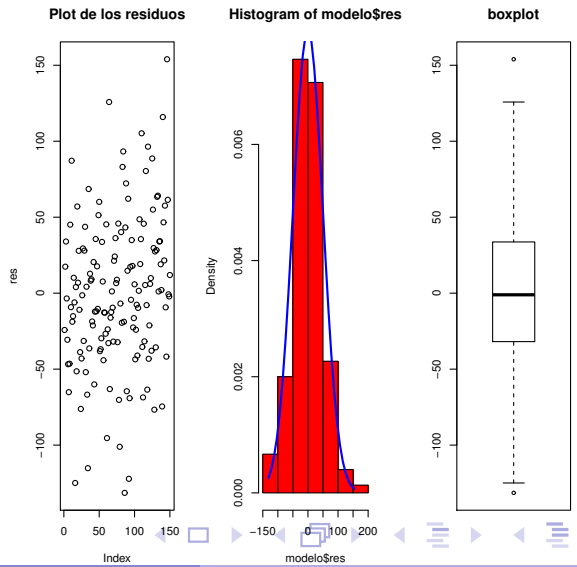
Un test de normalidad, por ejemplo Shapiro Wilks, KS, etc...

```
#Kolmogrov-Smirnov
>library(stats)
>ks.test(res/48.72,pnorm)
One-sample Kolmogorov-Smirnov test
data:  res/48.72
D = 0.054, p-value = 0.7752

> shapiro.test(res)
Shapiro-Wilk normality test
data: res
W = 0.9789, p-value = 0.7811
```

Acepto H0: variable normal

Un **histograma y boxplot de los residuos**:

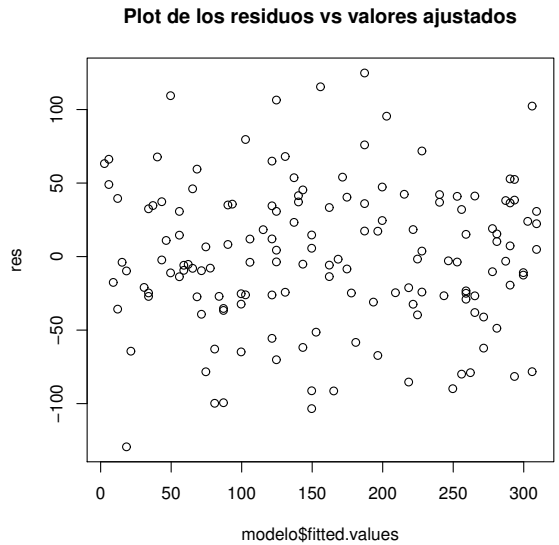
```
>par(mfrow=c(1,3))
>summary(modelo)
>modelo$res
>res=resid(modelo)
>plot(res,main=paste("Plot de los residuos"))
>hist(modelo$res,breaks=10,col="red",proba=T)
xfit=seq(min(res),max(res),length=31)
yfit=dnorm(xfit,mean=mean(res),sd=sd(res))
lines(xfit,yfit,col="blue",lwd=2)
>boxplot(modelo$res)
```



Gráficos asociados a errores

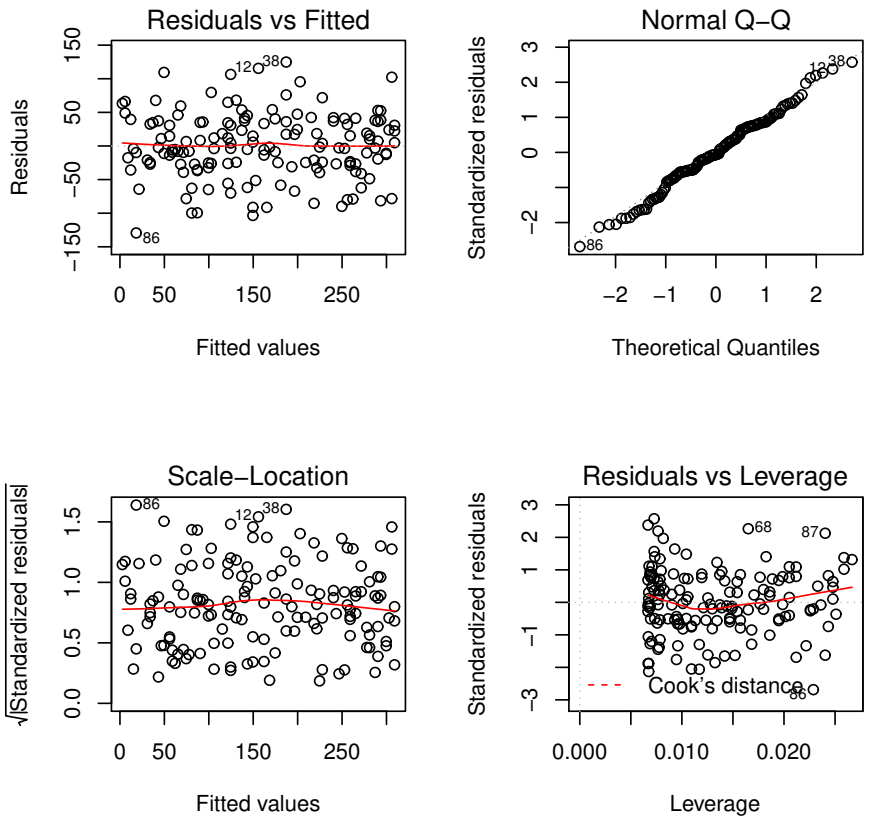
1. **Gráfico de dispersión de residuos respecto del índice**, con el fin de detectar agrupaciones o correlación contraria a la aleatoriedad.

2. **Gráfico de los residuos versus los valores ajustados**



Gráficos

plot(modelo)



1. El primer gráfico representa los residuos en función de los valores predecidos por el modelo. La linea roja (*smooth fit*) intenta mostrar una tendencia. En nuestro ejemplo, vemos que no hay tendencia.
2. El segundo gráfico es el qqplot. Con éste gráfico se visualiza el ajuste de la distribución muestral de los residuos a la ley normal estándar (por eso están estandarizados).

Distancia de Cook

Distancia de Cook

- La influencia de un punto i puede ser vista también comparando la estimación con o sin él.
- Se define la distancia de Cook de la i-esima observaci on como

$$C_i = \frac{(\widehat{y}_{(i)} - \widehat{y})'((\widehat{y}_{(i)} - \widehat{y}_i))}{2\widehat{\sigma^2}} = \frac{h_{ii} r_i^2}{2(1 - h_{ii})} \quad \forall\, i = 1, \ldots, n$$

  y entonces C_i es proporcional a la distancia euclidea entre $\widehat{y}_{(i)}$ e $\widehat{y}$.

- Si C_i es grande entonces la observación i tiene mucha influencia sobre $\widehat{\beta}$ y sobre $\widehat{y}$.
- La búsqueda de puntos influyentes se puede iniciar con la identificación de puntos con C_i elevada. Sin embargo, se desconoce la distribución de este estadístico y no hay regla para la determinación de puntos con valor de C_i grande. Algunos autores dicen que es significativa si es mayor a 1.

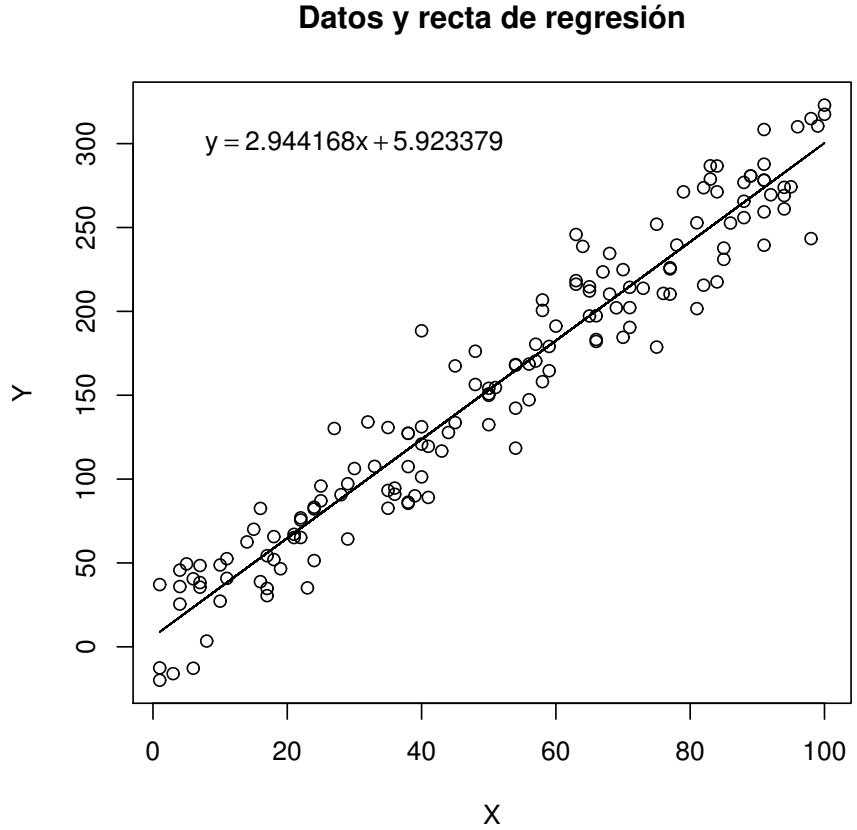
Plan

Ejemplo simulado

Simulemos 150 datos que provienen del modelo

$$Y = 2 + 3X + \epsilon \quad \epsilon \sim N(0, 50)$$

```
>x=1:100
>X=sample(x,150,replace=T)
>Y=2+3*X+rnorm(150,0,50)
>modelo=lm(Y~X)
> modelo

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
      5.923        2.944
```



```
> anova(modelo)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X           1 1097949 1097949  2269.6 < 2.2e-16 ***
Residuals 148   71598     484
```

```
> summary(modelo)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-51.064 -16.705   0.299  14.702  64.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9234     3.6476   1.624    0.107
X             2.9442     0.0618  47.640   <2e-16 ***
---

Residual standard error: 21.99 on 148 degrees of freedom
Multiple R-squared:  0.9388,    Adjusted R-squared:  0.9384
F-statistic:  2270 on 1 and 148 DF,  p-value: < 2.2e-16
```

Para ver los residuos y verificar supuesto de normalidad y de iid:

```
> modelo$res
```

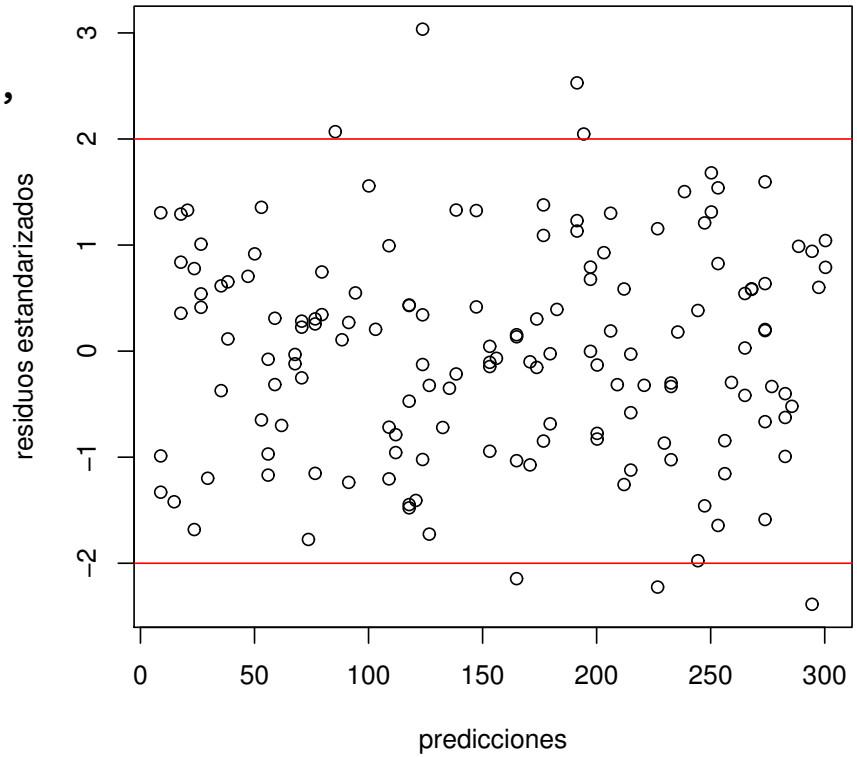
```
> rstudent(modelo)
```

Si un punto tiene residuo studentizado ($e_i/s.e(e_i)$) mayor que 2 en valor absoluto entonces el punto es sospechoso.

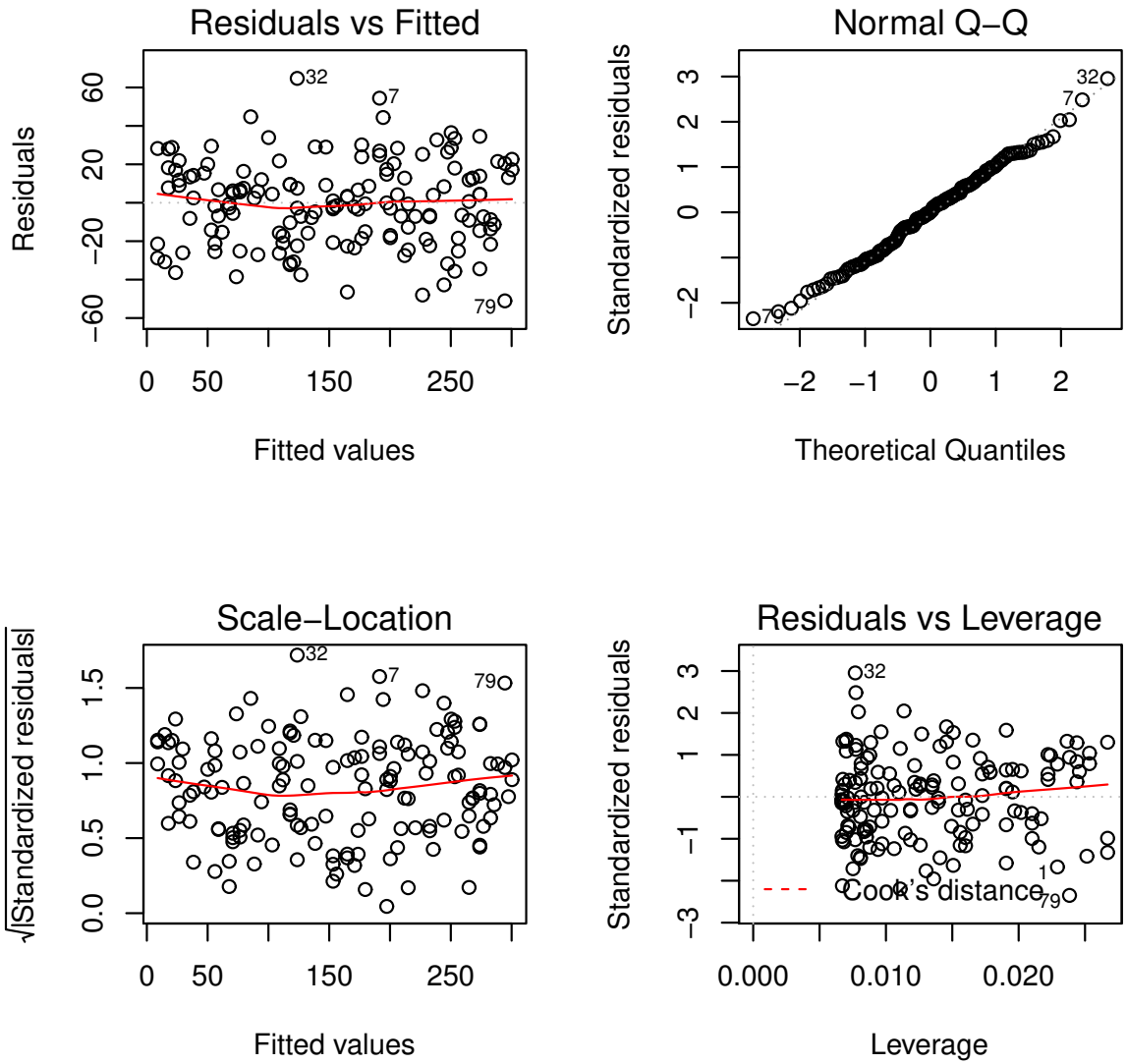
```
> plot(modelo$fitted,rstudent(modelo),
xlab="predicciones",
ylab="residuos estandarizados")
```

```
> abline(h=2,col="red")
```

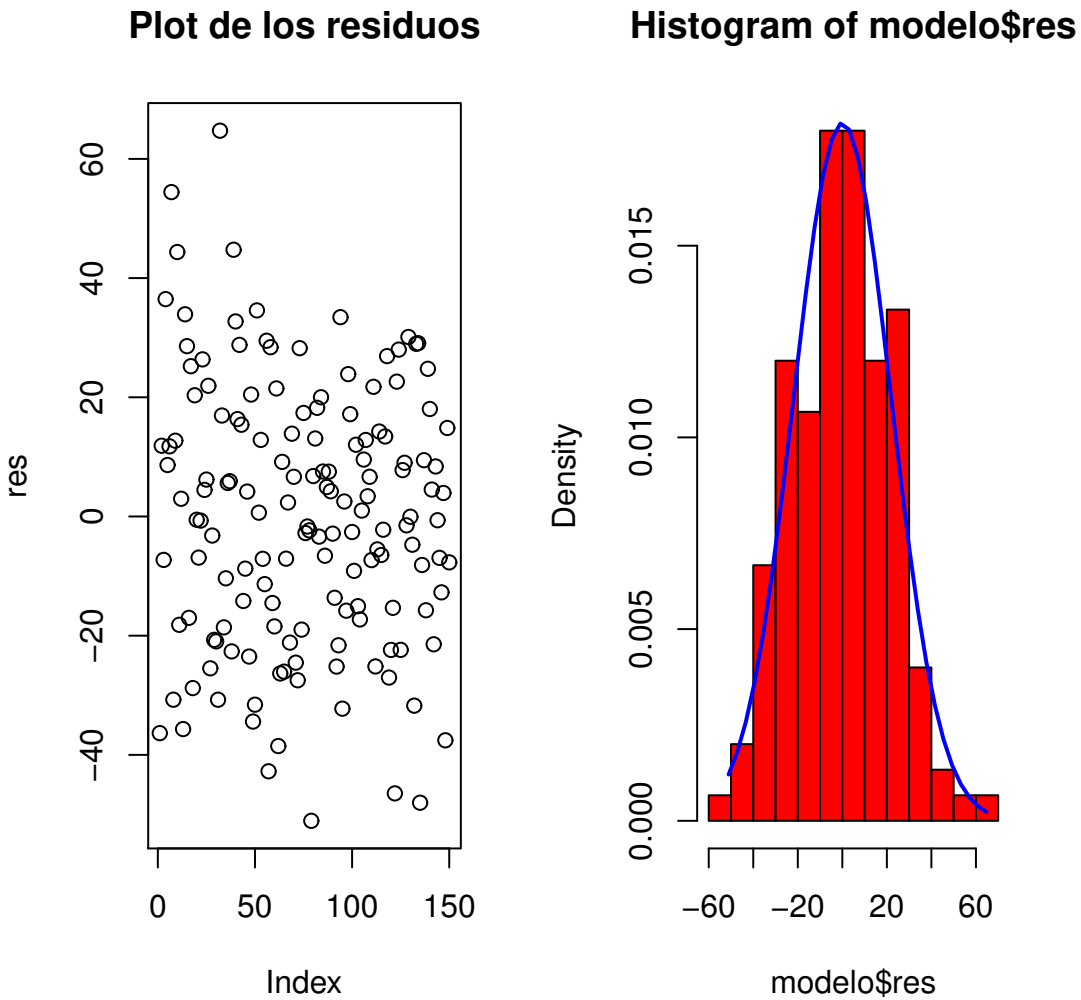
```
> abline(h=-2,col="red")
```



```
> par(mfrow=c(2,2))
> plot(modelo)
```



```
>res=resid(modelo)
>par(mfrow=c(1,2))
>plot(res,main=paste("Plot de los residuos"))
>hist(modelo$res,breaks=10,col="red",proba=T)
>xfit=seq(min(res),max(res),length=31)
>yfit=dnorm(xfit,mean=mean(res),sd=sd(res))
>lines(xfit,yfit,col="blue",lwd=2)
```



También se puede aplicar el test de Shapiro Wilks

```
> shapiro.test(res)

Shapiro-Wilk normality test

data: res

W = 0.9789, p-value = 0.7811
```

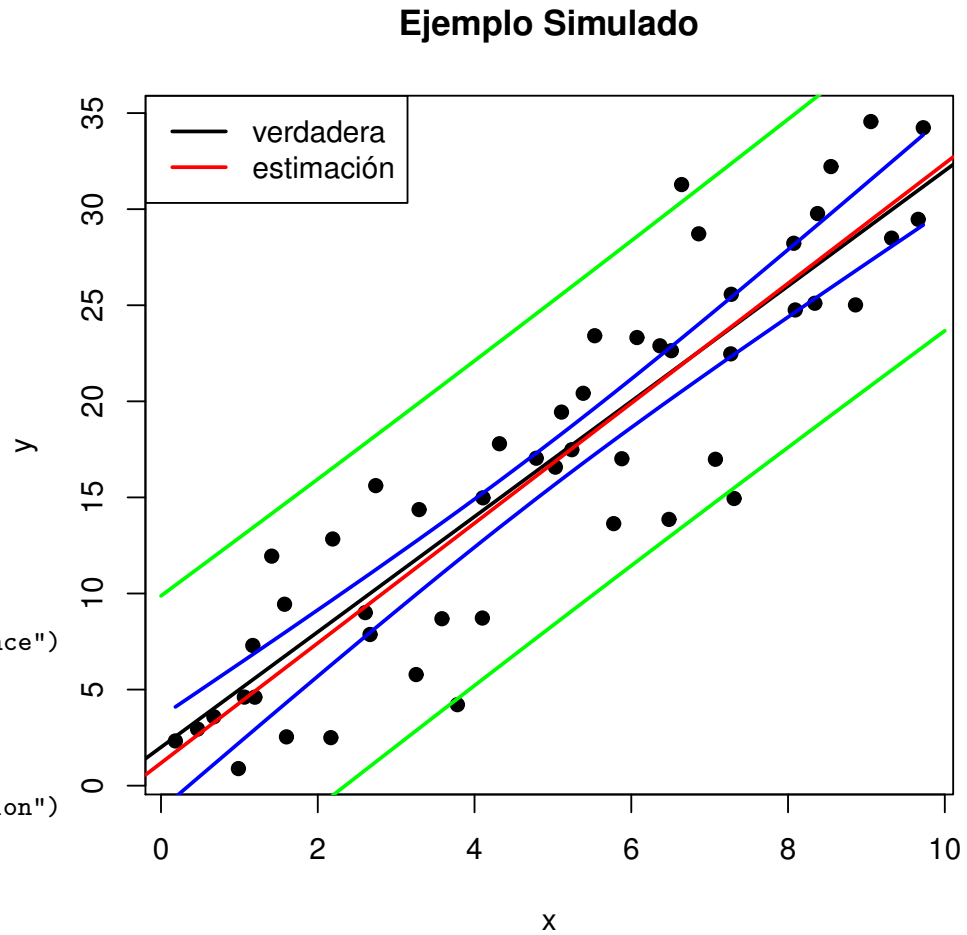
Acepto H0: variable normal

Interpretación y conclusión sobre el modelo

1. los residuos parecerían ser gaussianos e indenticamente distribuidos.
2. El modelo tiene una buena performance explicativa $R^2 = 0.9388$ (cerca de 1) y el error residual (residual standard error, RSE), $\widehat{\sigma} = \sqrt{\frac{SCR}{n-2}}$, es bajo $(21, 99)$ por lo que augura buenas predicciones.
3. los errores estandares de $\widehat{\beta}_0$ (3.64) y $\widehat{\beta}_1$ (0.06) son pequeños: esto indica una cierta estabilidad del modelo.
4. El termino constante no es significativamente distinto de cero (podríamos prescindir de él).
5. El coeficiente en X, β_1, es significativamente distinto de cero.
   Otra manera de verlo: el $F = 2269.6$. Hay fuerte evidencia de que $\beta_1 \neq 0$.

Intervalo de confianza para la recta de regresión e intervalo de confianza de una predicción

```
n <- 50
x <- sort(10 * runif(n))
y <- 2 + 3 * x + rnorm(n, sd = 5)
fit <- lm(y ~ x)
plot(x, y, pch = 19)  # datos
abline(2, 3, lwd = 2) # verdadera
abline(coef(fit), lwd = 2, col = 'red') # estimaci\'on
legend("topleft", c("verdadera", "estimaci\'on"),
lty = 1, lwd = 2, col = c(1, 2))

new=data.frame(x=seq(0, 10, .5))
pred=predict(fit, interval="confidence")
pred2=predict(fit,newdata=new,interval="prediction")
lines(x, pred[, 2], col = "blue", lwd = 2)
lines(x, pred[, 3], col = "blue", lwd = 2)
lines(new[,1], pred2[, 2], col = "green", lwd = 2)
lines(new[,1], pred2[, 3], col = "green", lwd = 2)
title("Ejemplo Simulado")

>predict(fit,newdata=data.frame(x=c(5,6)),interval="confidence")
       fit      lwr      upr
1 16.77854 15.59662 17.96046
2 19.89853 18.63624 21.16082

>predict(fit,newdata=data.frame(x=c(5,6)),interval="prediction")
       fit       lwr      upr
1 16.77854  8.338984 25.21810
2 19.89853 11.447340 28.34972
```



Referencias

- A. I. Izenman, *Modern Multivariate Statistical Techniques*, Springer, 2008.
- F. Carmona, *Modelos Lineales*, notas de curso, Universitat de Barcelona, 2003.
- C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006
- S.J. Sheater, *A Modern Approach to Regression with R*, Springer, 2009.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.
- M. Bourel. Apuntes curso Estadística Multivariada Computacional 2018, 2019. Facultad de Ingeniería.