Probabilidad y Estadística 2017
Una introducción a la regresión lineal simple y múltiple

Mathias Bourel

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

2 de noviembre de 2017

Plan

1. Regresión lineal simple.
2. Regresión lineal múltiple
3. Pruebas de hipótesis

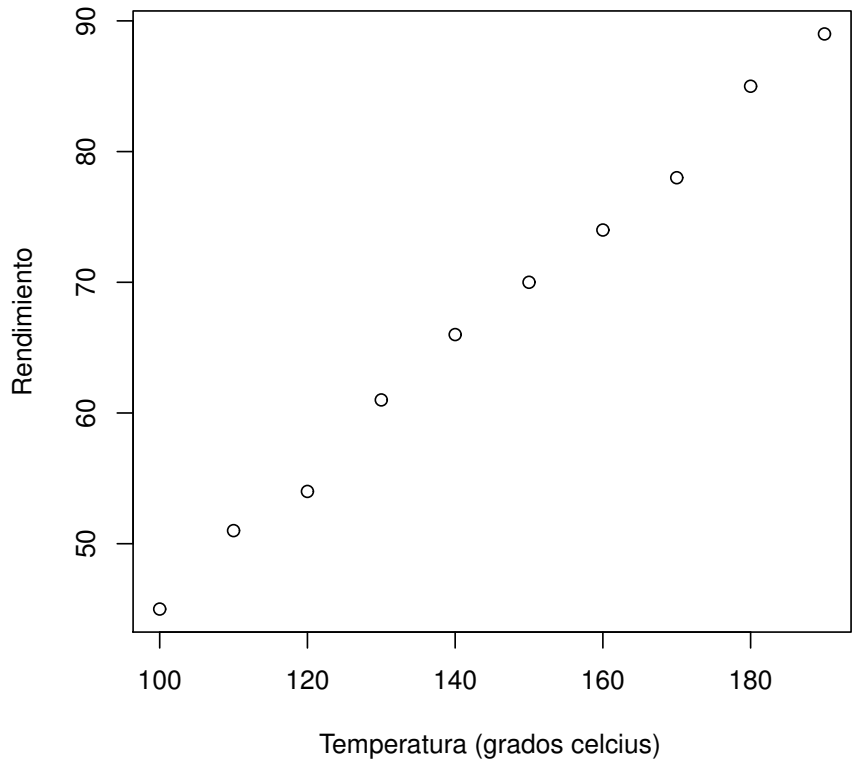
Plan

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.

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:

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

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.

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

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

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

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

Método de los mínimos cuadrados

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

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

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

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.

Historia

TABLE XXII.

Father's Stature and Son's Stature.

Father's Stature.

Son'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

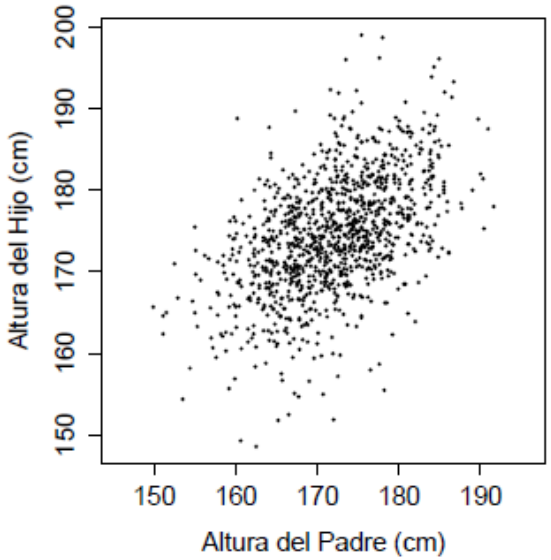
53—2

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

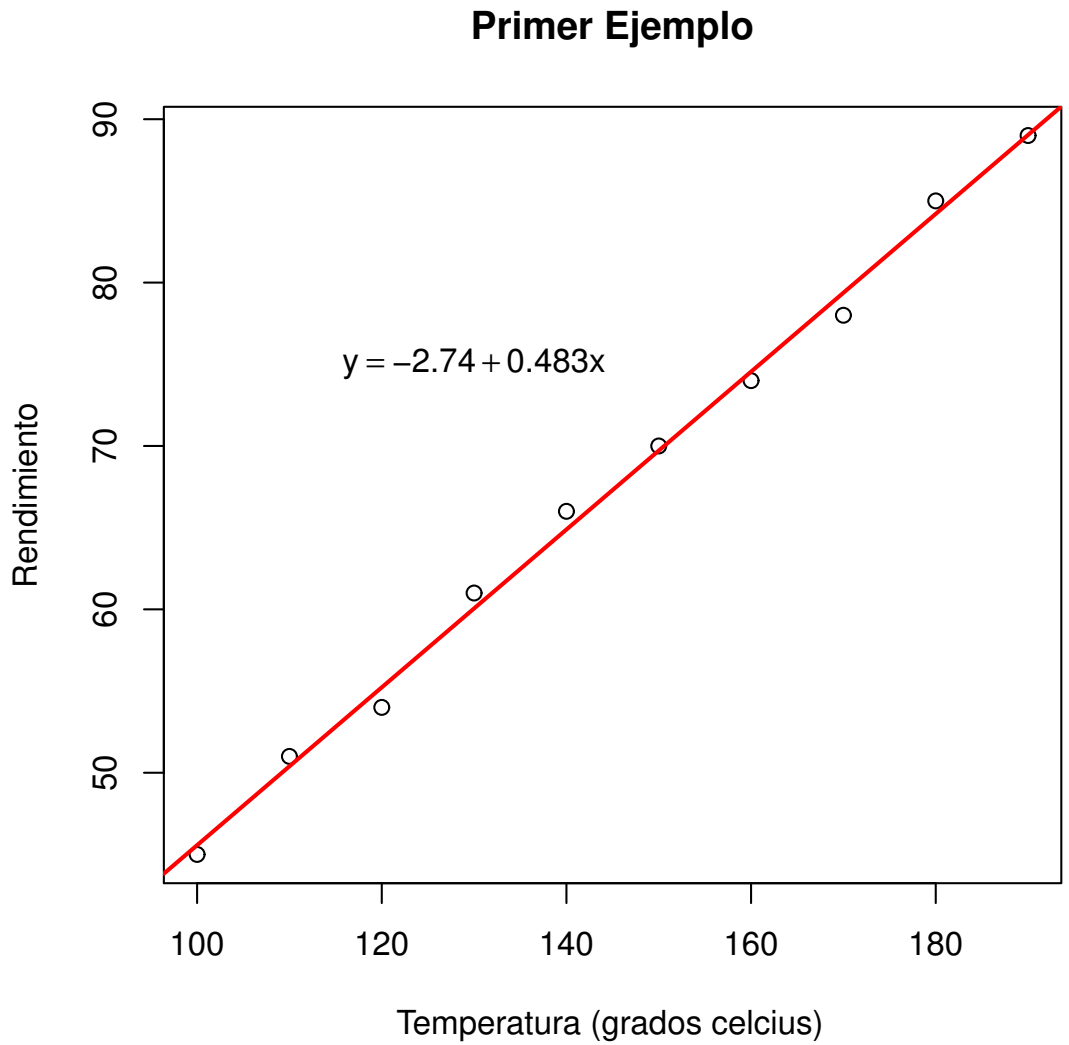
Volvemos a nuestro problema inicial:

```
>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.

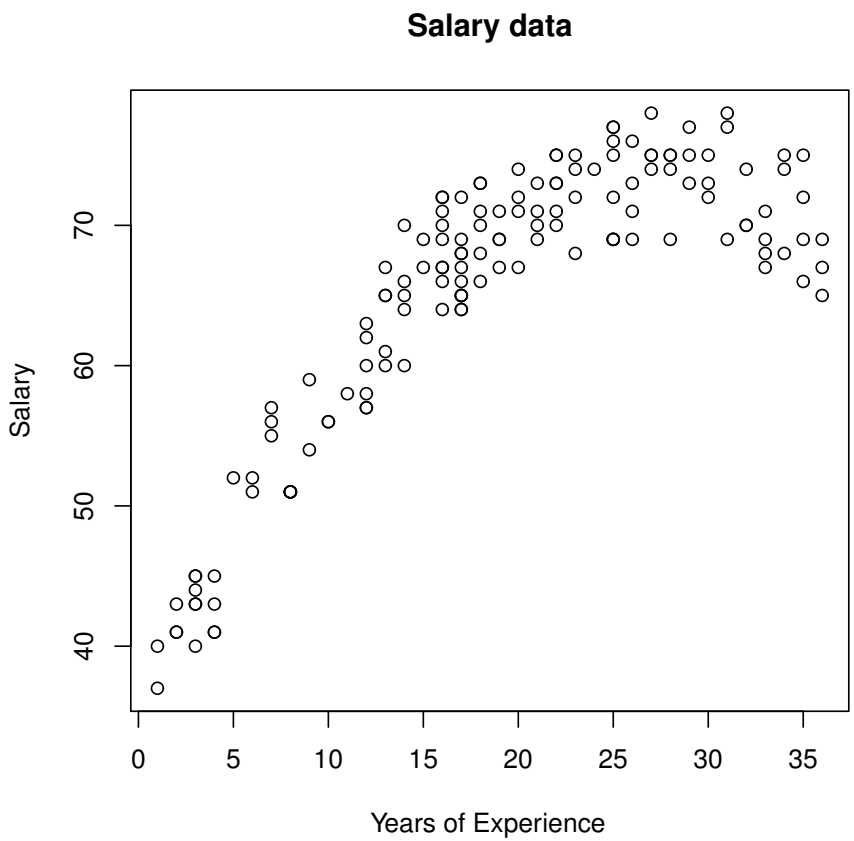
Con el R, las funciones que se usan son:

```
>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

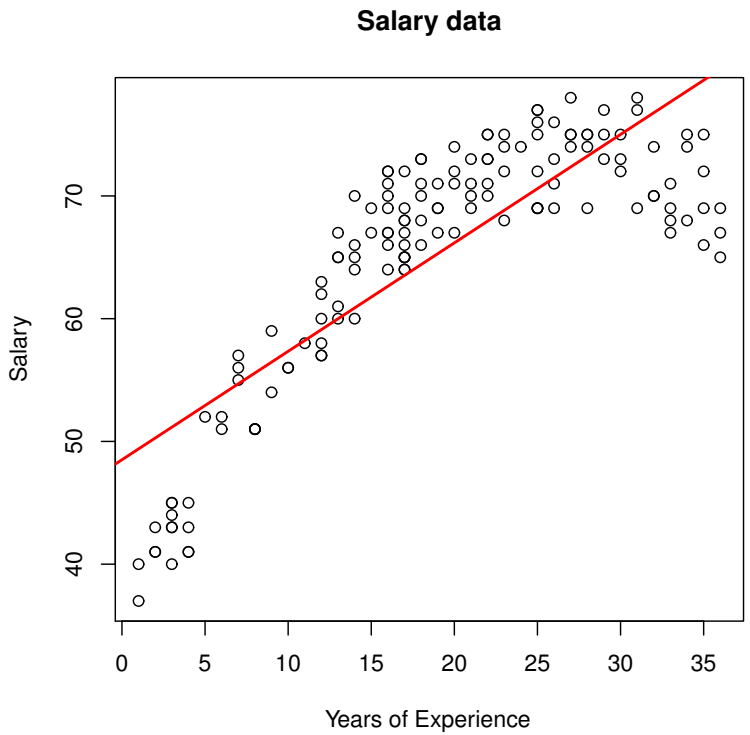
Claramente esta relación no es lineal y no sería adecuada el modelo de regresión lineal simple

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

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)
```

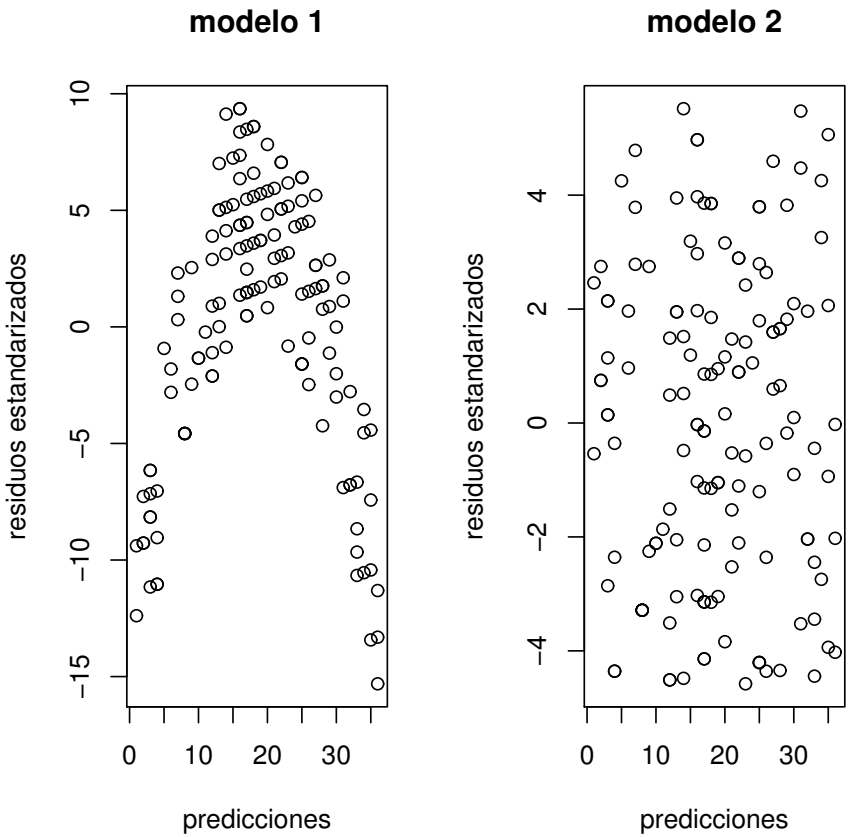



```
>m2 <- lm(Salary~Experience +
I(Experience^2))
```

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

>
```

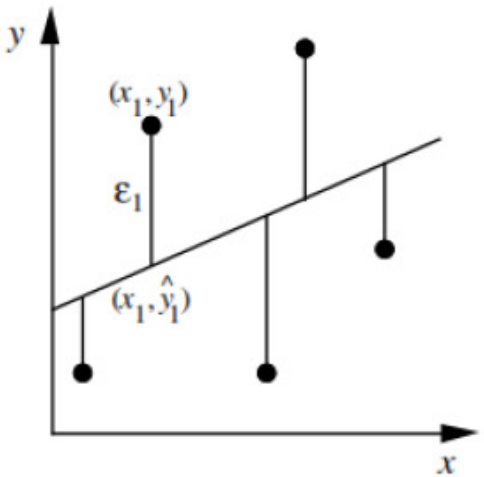
Plan

Generalización

Ecuación fundamental:

$$\text{“observación”} = \text{“modelo”} + \text{“error aleatorio”}$$
$$Y = F(X) + \epsilon$$

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

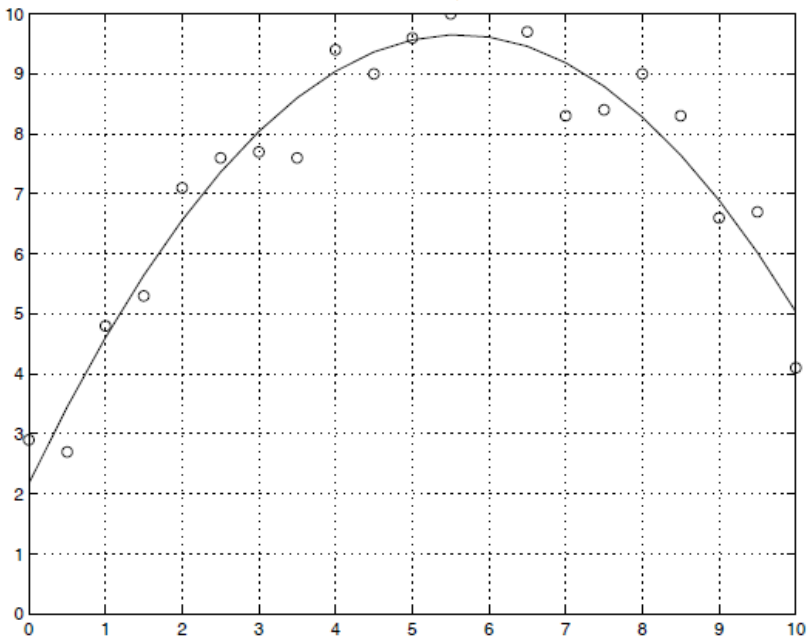
$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n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

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!)

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

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

El estimador mínimos cuadrados se basa en la siguiente propiedad de la proyección ortgonal: si $v \in \mathbb{R}^{d+1}$ y S es un subespacio de $\mathbb{R}^{d+1}$ entonces

$$||v - P_S(v)|| \leq ||v - s|| \ \ \forall s \in S$$

Acá consideramos como subespacio S al subespacio generado por las columnas de X y notaremos $S = \langle X \rangle \in \mathbb{R}^n$.

Observar que el complemento ortogonal de S es

$$S^{\perp} = \langle X \rangle^{\perp} = \{v \in \mathbb{R}^n : X'v = 0_{\mathbb{R}^{d+1}}\} = N(X')$$

Prueba: $v \in \langle X \rangle^{\perp} \Leftrightarrow v$ es ortogonal a todas las columnas de $X \Leftrightarrow v$ es ortogonal a todas las filas de X'
$\Leftrightarrow X'v = 0_{\mathbb{R}^{d+1}} \Leftrightarrow v \in N(X')$

Buscamos entonces $\widehat{\beta} \in \mathbb{R}^{d+1}$ tal que $X\widehat{\beta} = P_S(Y)$. Entonces:

$$Y - X\widehat{\beta} = Y - P_S(Y) = P_{S^{\perp}}(Y)$$

Entonces

$$X'P_{S^{\perp}}(Y) = 0_{\mathbb{R}^{d+1}} \Leftrightarrow X'\left(Y - X\widehat{\beta}\right) = 0_{\mathbb{R}^{d+1}} \Leftrightarrow X'Y = X'X\widehat{\beta}$$

Generalización

A la expresión $X'Y = X'X\widehat{\beta}$ se le llama *ecuaciones normales*. 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

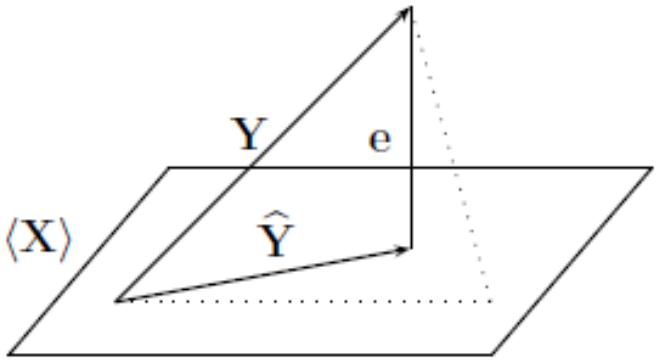
$$\widehat{\beta} = (X'X)^{-}X'Y$$

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

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



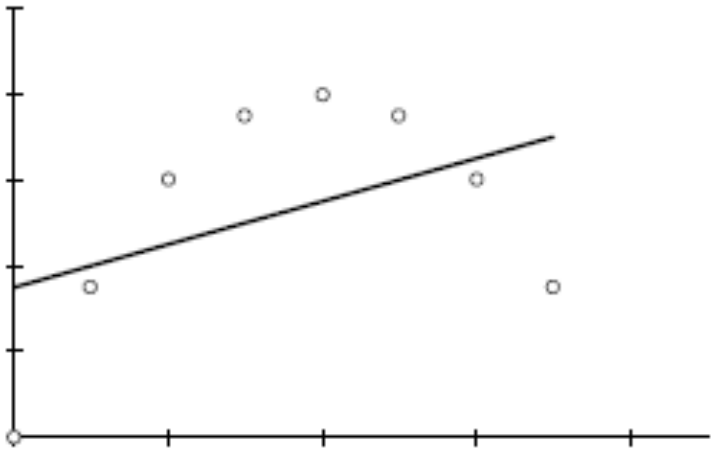
Entonces

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

Supuestos: condiciones de Gauss-Markov

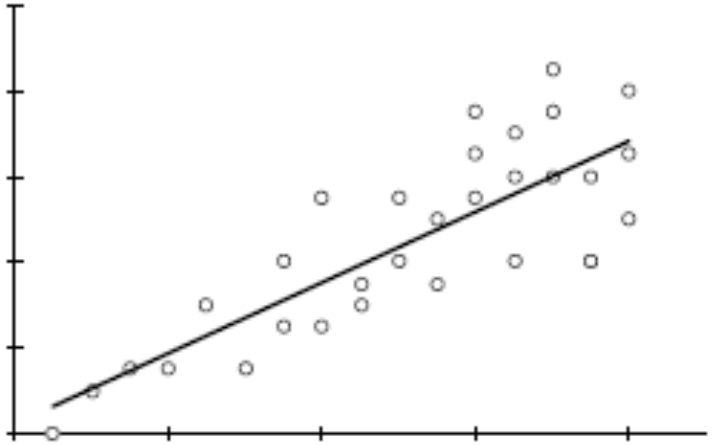
Volvemos a la regresión lineal simple. Hasta ahora el método de los mínimos cuadrados es análitico. Veamos donde interviene la estadística.
Suponemos que $x_1, \ldots, x_n$ son constantes. Supongamos que los errores e_i provienen de una variable aleatoria ϵ e impondremos que estos errores verifiquen 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):



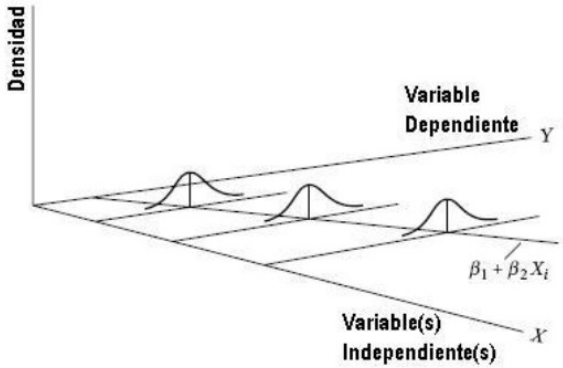


Figura: homocedasticidad

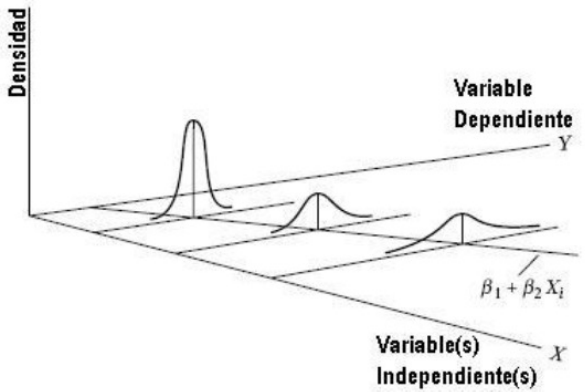


Figura: heterocedasticidad

Supuestos: condiciones de Gauss-Markov

(3) Las observaciones deben ser incorrelacionadas.

Todo esto se puede resumir como

$$\mathbb{E}(\epsilon) = \mathbf{0}_{\mathbb{R}^n} \qquad Var(\epsilon) = \sigma^2 I_n$$

donde $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ y $Var(\epsilon)$ es la matriz de varianza-covarianza de ϵ.

Resumen general

En resumen:
La expresión general del modelo lineal es:

$$Y = \underbrace{X\beta}_{f(X)} + \epsilon$$

y la estimación:

$$\widehat{Y} = X\widehat{\beta}$$

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 $Y = X\beta + \epsilon$ cumple que

$$\mathbb{E}(Y) = X\beta$$

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:

$$Y \sim N_n(X\beta, \sigma^2 I_n)$$

Ejemplo: regresión lineal simple

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 $\sum(\widehat{y_i} - \overline{y})^2 = \widehat{\beta}_1^2 S_{x_i}$

Ejemplo: regresión lineal simple

Por otro lado: $Var(\widehat{\beta}) = \begin{pmatrix} var(\widehat{\beta}_0) & cov(\widehat{\beta}_0, \widehat{\beta}_1) \\ cov(\widehat{\beta}_0, \widehat{\beta}_1) & var(\widehat{\beta}_1) \end{pmatrix} = \sigma^2 (X'X)^{-1}$

Identificando se tiene:

$$\begin{pmatrix} var(\widehat{\beta}_0) & cov(\widehat{\beta}_0, \widehat{\beta}_1) \\ cov(\widehat{\beta}_0, \widehat{\beta}_1) & var(\widehat{\beta}_1) \end{pmatrix} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\overline{x}^2}{S_x} & -\frac{\overline{x}}{S_x} \\ -\frac{\overline{x}}{S_x} & \frac{1}{S_x} \end{pmatrix}$$

donde $\overline{x} = \frac{1}{n}\sum x_i$ y $\overline{y} = \frac{1}{n}\sum y_i$ (medias muestrales);

$s_x^2 = \frac{1}{n}\sum (x_i - \overline{x})^2$ y $s_y^2 = \frac{1}{n}\sum (y_i - \overline{y})^2$ (varianzas muestrales);

$S_x = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = \sum (x_i - \overline{x})^2 = n s_x^2$.

Entonces:

1. $\mathbb{E}(\widehat{\beta}_0) = \beta_0$ y
$$s.e(\widehat{\beta}_0)^2 = var(\widehat{\beta}_0) = \widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_x} \right) = \frac{SCR}{n-2} \left(\frac{1}{n} + \frac{\overline{x}^2}{S_x} \right)$$
2. $\mathbb{E}(\widehat{\beta}_1) = \beta_1$ y
$$s.e(\widehat{\beta}_1)^2 = var(\widehat{\beta}_1) = \widehat{\sigma}^2 \frac{1}{S_x} = \frac{SCR}{n-2} \frac{1}{S_x}$$

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:*

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

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

Entonces:

1. La variación total es $VT = \sum_{i=1}^{n}(y_i-\overline{y})^2 = S_y$
2. La variación no explicada por la regresión es $VNE = \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 = SCR$
3. La variación explicada por la regresión es $VE = \sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2 = \widehat{\beta}_1^2 S_x = SSR$

Regresión lineal simple. Descomposición variación

De la descomposición de la variación tenemos que:

$$S_y = SCR + \underbrace{\widehat{\beta}_1^2 S_x}_{SSR}.$$

Más aún:

$$SCR = (1 - r^2)S_y$$

donde r es el coeficiente de correlación muestral entre x e y.

$$SCR = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = S_y - \widehat{\beta}_1^2 S_x = S_y - \frac{S_{xy}^2}{S_x^2} S_x = S_y - \frac{S_{xy}^2}{S_x} = S_y - r^2 S_y = (1 - r^2)S_y$$

De la cuenta anterior tenemos

$$S_y = SCR + \widehat{\beta}_1^2 S_x = SCR + r^2 S_y$$

Por otro lado una estimación de σ^2 es

$$\widehat{\sigma}^2 = \frac{(1 - r^2)S_y}{n - 2}$$

Regresión lineal simple. Coeficiente de determinación R^2

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

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

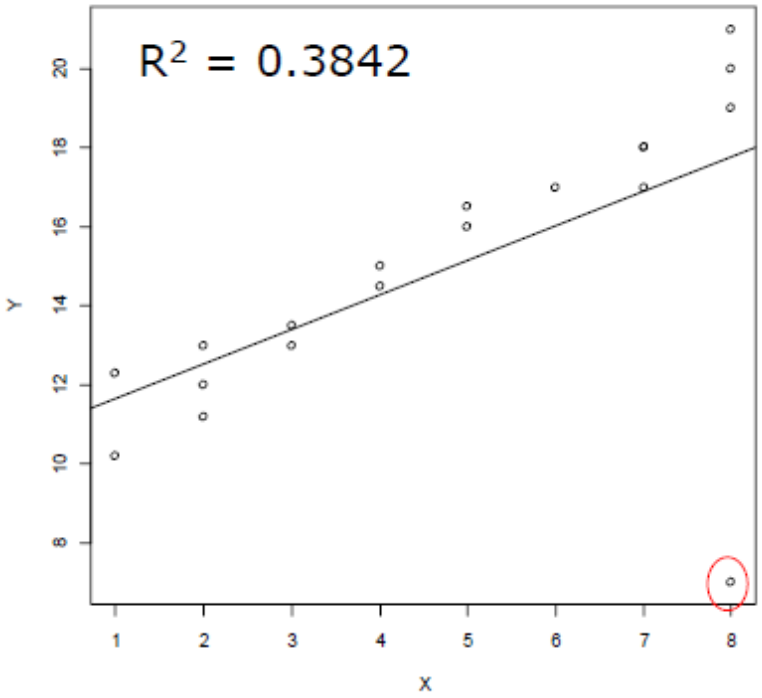
El coeficiente de determinación R^2 es una medida de la bondad del ajuste, **suponiendo que el modelo es lineal**. *En el caso de la regresión lineal simple* coincide con r^2.

- 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

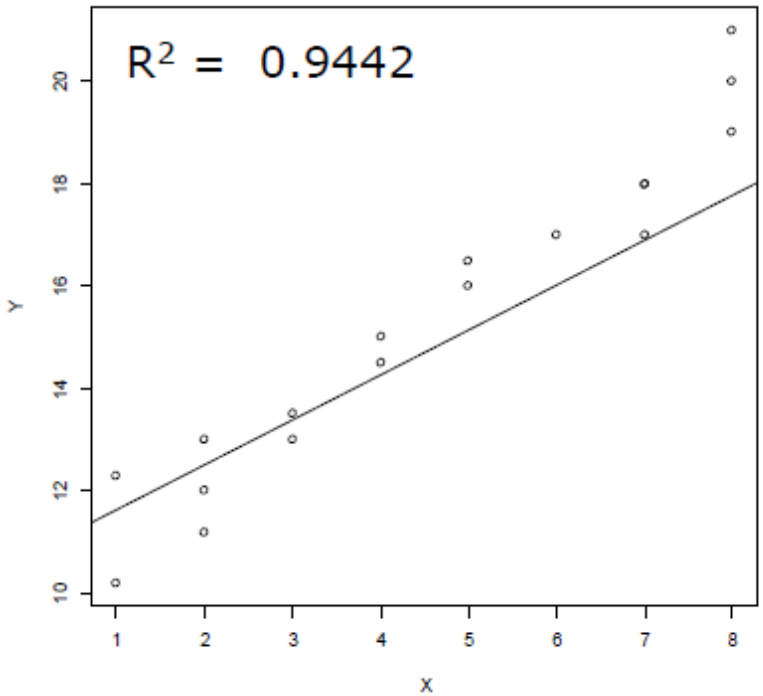
$$\overline{R}^2 = 1 - \frac{SCR/(n-2)}{S_y/(n-1)}$$

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



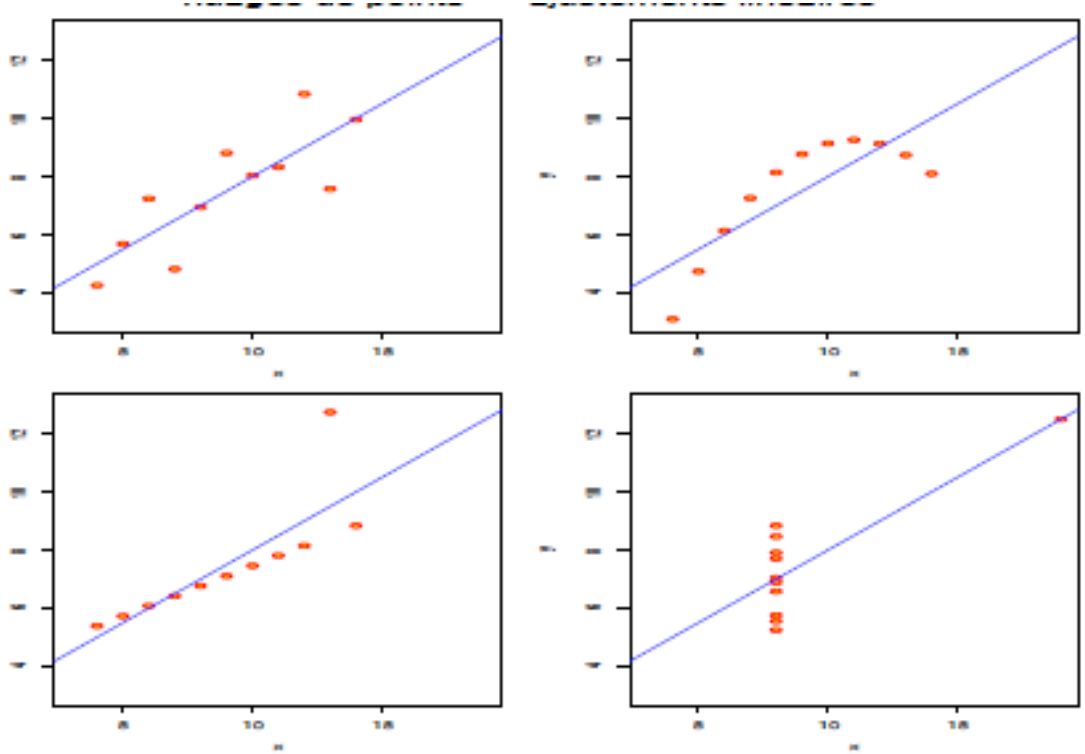
Comportamiento del R^2 con y sin un dato «outlier» en la variable Y.



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

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 $F = MSM/MSE$ 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**:**

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

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:

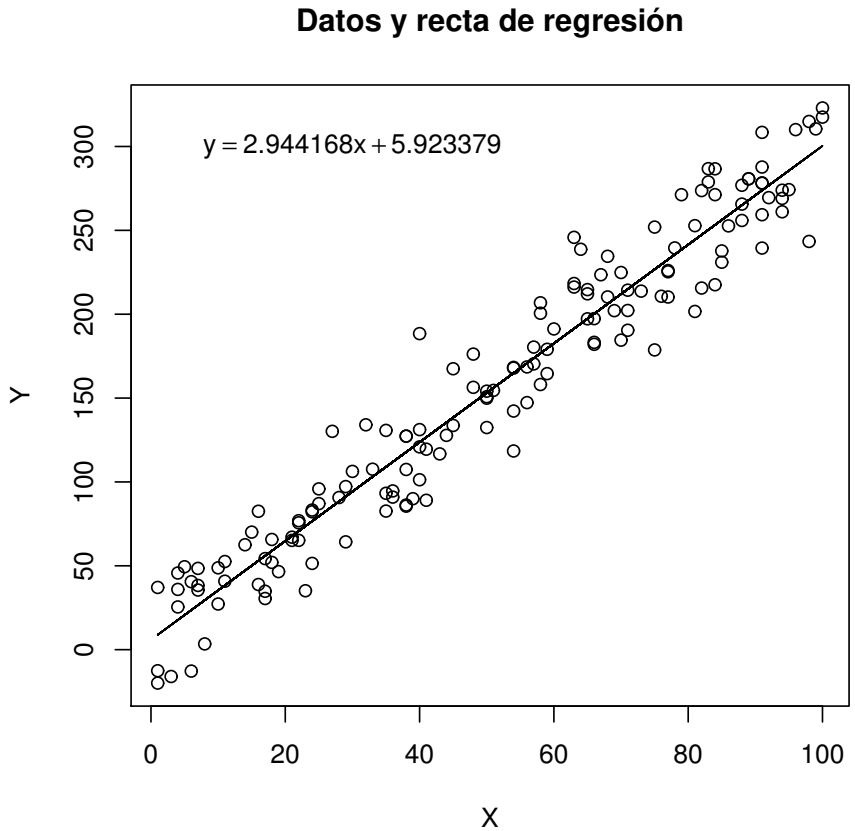
$$\left[\frac{SCR}{\chi^2_{n-2}(\alpha/2)}, \frac{SCR}{\chi^2_{n-2}(1-\alpha/2)}\right]$$

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
```

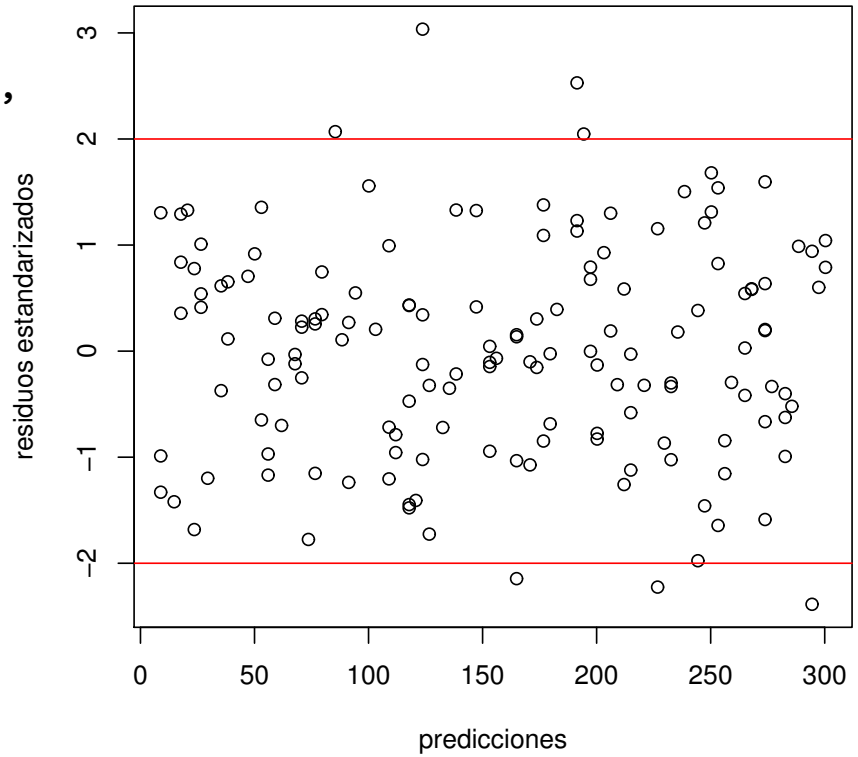
```
> rstudent(modelo)
```

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

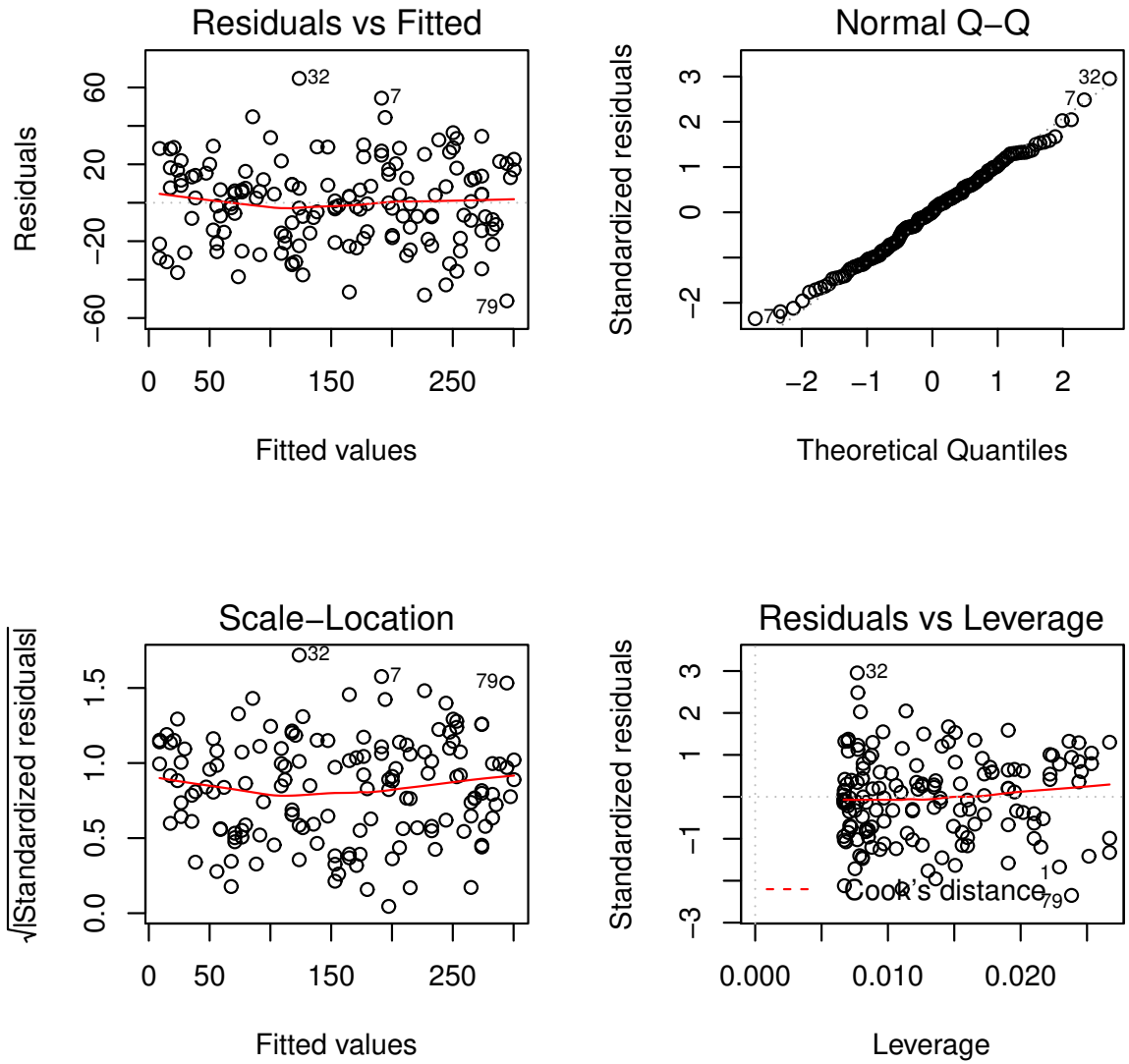
```
> plot(modelo$fitted,rstudent(modelo),
xlab="predicciones",
ylab="residuos estandarizados")
```

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

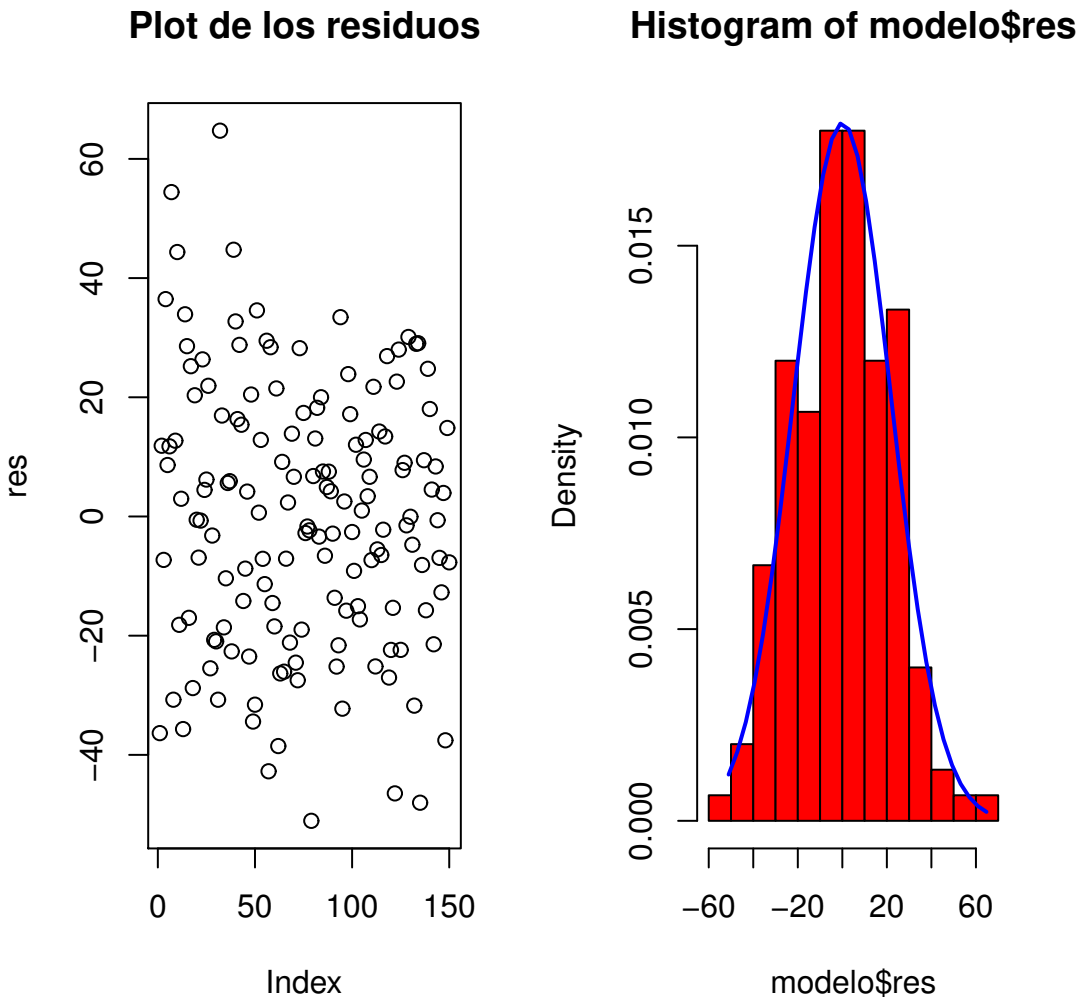
```
> abline(h=-2,col="red")
```



```
> 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 $\beta_1 \neq 0$.

En este test de hipótesis, bajo hipotesis normalidad, nos preguntamos si los coeficientes de la regresión lineal son nulos o no. Es el análogo al test t de la regresión lineal simple.

En regresión lineal múltiple:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0 \\ H_1 : \text{al menos un } \beta_j \text{ es no nulo} \end{cases}$$

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

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 2016, 2017. Facultad de Ingeniería.