Comprobación de las hipótesis del modelo: diagnosis

Contents

1 Hipótesis del modelo y residuos **1**

2 Análisis gráfico de los residuos **2**
2.1 Residuos frente a valores predichos 2
2.2 Gráfico Cuantil-Cuantil 4
2.3 Gráfico de residuos estandarizados vs valores predichos. 6
2.4 En R 7

3 Datos atípicos **8**

1 Hipótesis del modelo y residuos

Recordamos el modelo:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i,\ i = 1, 2, \cdots, n$$

Además se considera que $u_i \sim N(0, \sigma^2)$. Esta hipótesis implica:

- Normalidad
- Varianza constante: todos los u_i tienen la misma varianza.
- Independepencia: $Cov[u_i, u_j] = 0\ \forall i, j,\ i \neq j$

La variable aleatoria u_i no se observa, por lo que se trabaja con los residuos:

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}) = y_i - \hat{y}_i$$

En forma vectorial:

$$e = y - \hat{y} = (I - H)y$$

ya que $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$, donde $H = X(X^T X)^{-1} X^T$. Sustituyendo el valor de y:

$$e = (I - H)(X\beta + U) = (I - H)U$$

ya que $HX = X$. Por tanto, los errores del modelo y los residuos no son intercambiables, sino que los residuos son una combinación lineal de los errores.

Debido a que $U \sim N(0, \sigma^2 I)$, los residuos tienen distribución $e \sim N(0, \sigma^2(I - H))$ puesto que

$$E[e] = 0$$

$$Var(e) = \sigma^2(I - H)$$

ya que la matriz H es simétrica ($H = H^T$) e idempotente ($H \cdot H = H$). La varianza de cada residuo viene dada por:

$$Var(e_i) = \sigma^2(1 - h_{ii}), \quad i = 1, \ldots, n$$

donde h_{ii} es el elemento de la diagonal de H. A partir de estos valores se definen los *residuos estandarizados*:

$$r_i = \frac{e_i}{\hat{s}_R\sqrt{1 - h_{ii}}}$$

que tienen varianza aproximadamente uno ya que la varianza residual es un estimador centrado de la varianza del modelo.

Los residuos estandarizados se utilizan para comprobar las hipótesis del modelo.

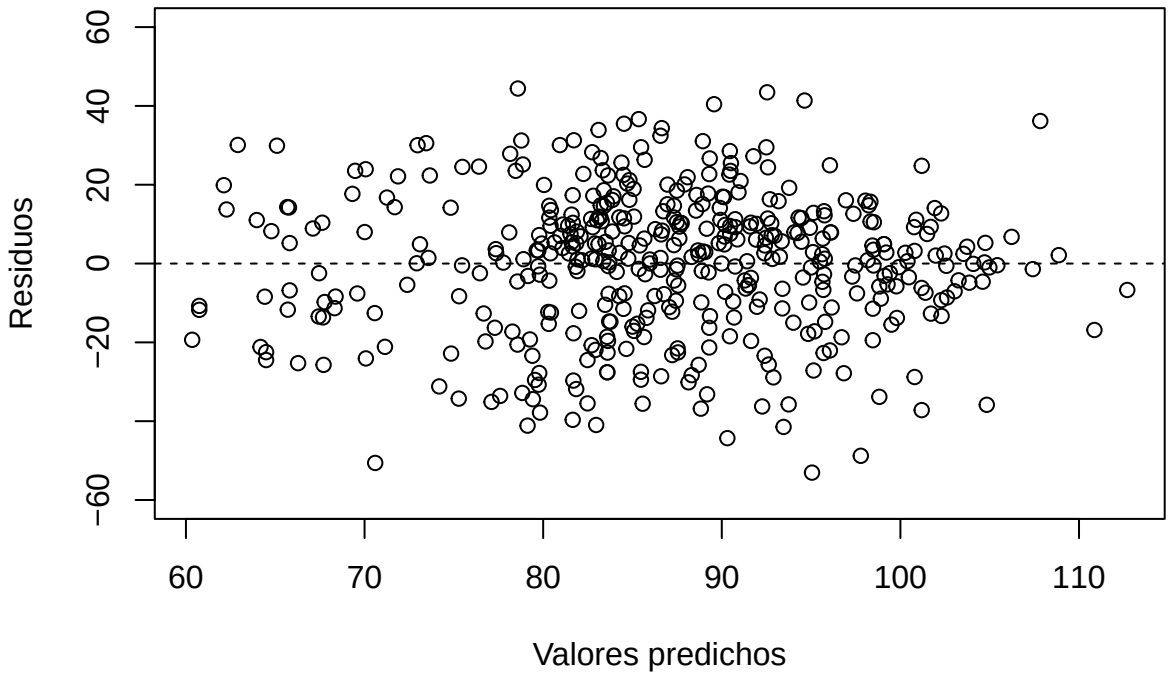
2 Análisis gráfico de los residuos

En general se prefiere analizar los residuos de manera gráfica en lugar de un análisis más cuantitativo.

2.1 Residuos frente a valores predichos

La herramienta más usada es el gráfico de residuos frente a valores predichos

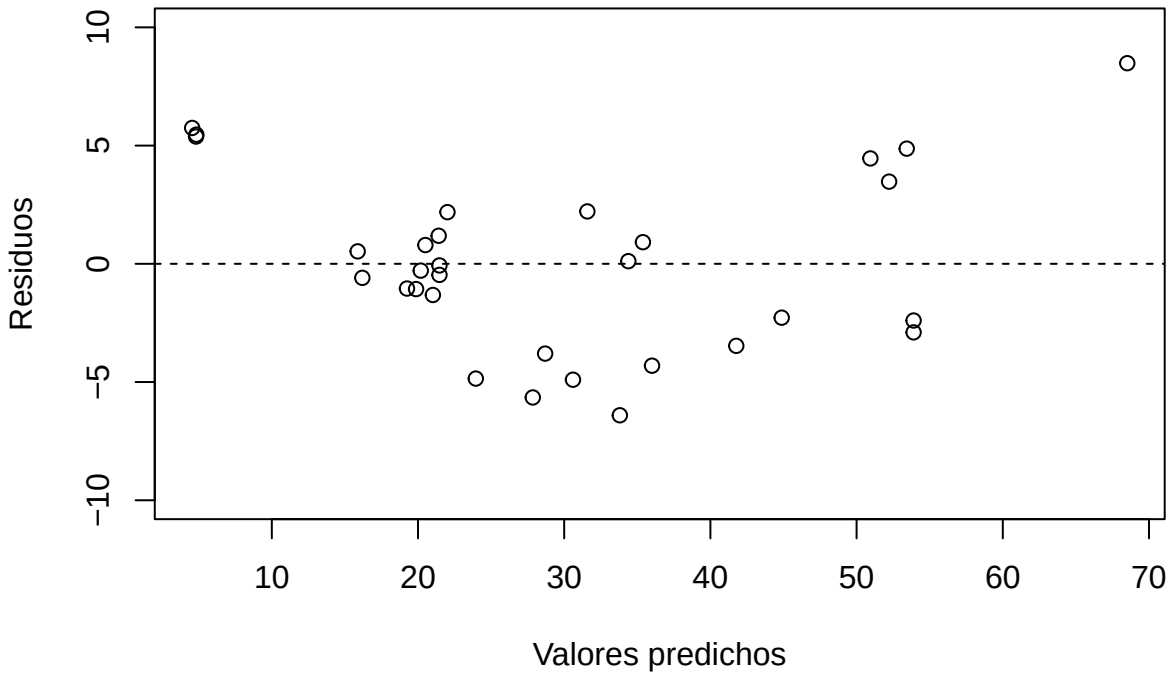
```
d = read.csv("datos/kidiq.csv")
d$mom_hs = factor(d$mom_hs, labels = c("no","yes"))
d$mom_work = factor(d$mom_work)
m = lm(kid_score ~ mom_iq * mom_hs + mom_age + mom_work, data = d)
#
plot(fitted(m),residuals(m), xlab = "Valores predichos", ylab = "Residuos", ylim = c(-60,60))
abline(h=0, lty = 2)
```



- Este gráfico se utiliza para comprobar que la varianza de los residuos es constante (también conocido como homocedasticidad) y que el modelo es lineal.

- Los residuos se deben distribuir homogéneamente a un lado y otro del eje X.
- Cuando hay heterocedasticidad la dispersión de los residuos suele aumentar con el valor de $\hat{y}_i$.
- Cuando no hay linealidad se observa curvatura en los residuos.
- Por ejemplo, sean los datos:

```
d2 = read.table("datos/cerezos.txt", header = T)
m2 = lm(volumen ~ diametro + altura, data = d2)
plot(fitted(m2),residuals(m2), xlab = "Valores predichos", ylab = "Residuos", ylim = c(-10,10))
abline(h=0, lty = 2)
```

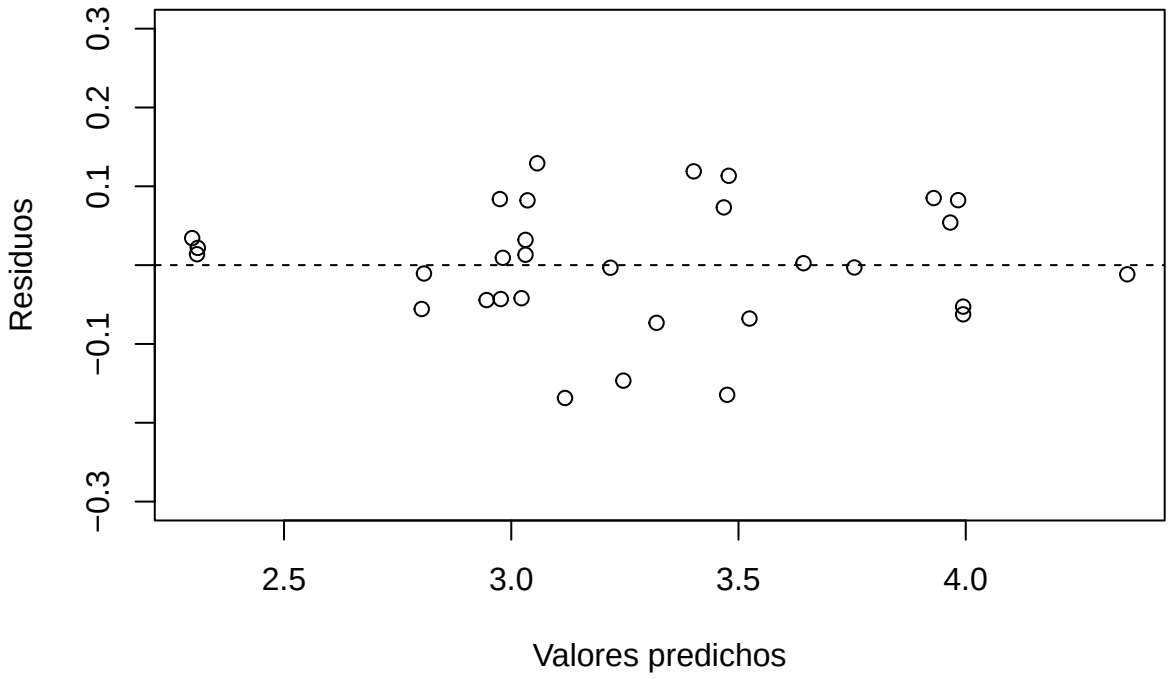


En este caso, no se cumple homocedasticidad ni linealidad.

- Los problemas de falta de linealidad se pueden corregir con transformaciones.
- Los problemas de varianza no constante se pueden corregir también con transformaciones.
- Las transformaciones puede ser de una variable, de varias variables o de todas las variables simultaneamente.
- Las transformaciones más usuales son: sqrt, log, x^2, 1/x
- También se puede utilizar minimos cuadrados generalizados para corregir los problemas de varianza.

En este caso, utilizamos transformaciones de todas las variables:

```
m2a = lm(log(volumen) ~ log(diametro) + log(altura), data = d2)
plot(fitted(m2a),residuals(m2a), xlab = "Valores predichos", ylab = "Residuos", ylim = c(-0.3,0.3))
abline(h=0, lty = 2)
```



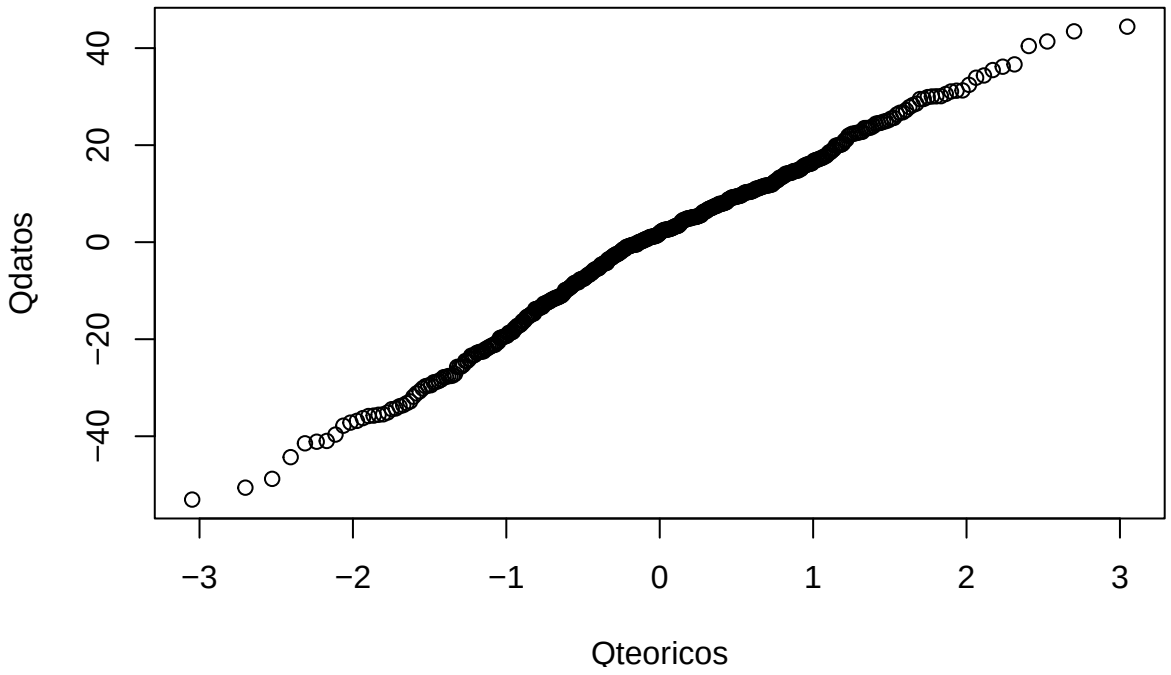
2.2 Gráfico Cuantil-Cuantil

Se utiliza para verificar la hipótesis de normalidad. En este gráfico se comparan los n datos ordenados de menor a mayor con n datos ordenados que provienen de la distribución N(0,1). Para obtener estos datos teóricos se consideran los cuantiles de la N(0,1), es decir, los puntos tales x_i que a $P(X \leq x_i) = (i - 0.5)/n$, donde n es el número de datos, $i = 1, 2, \cdots, n$. De esta manera obtenemos valores teóricos distribuimos uniformemente en el rando de la variable N(0,1). En realidad dichos valores x_i son cuantiles de la N(0,1).

```
Qdatos = sort(residuals(m))
#
n = length(residuals(m))
i = 1:n
q = (i-0.5)/n
Qteoricos = qnorm(q)
```

- Se representa Qteoria *vs* Qdatos.

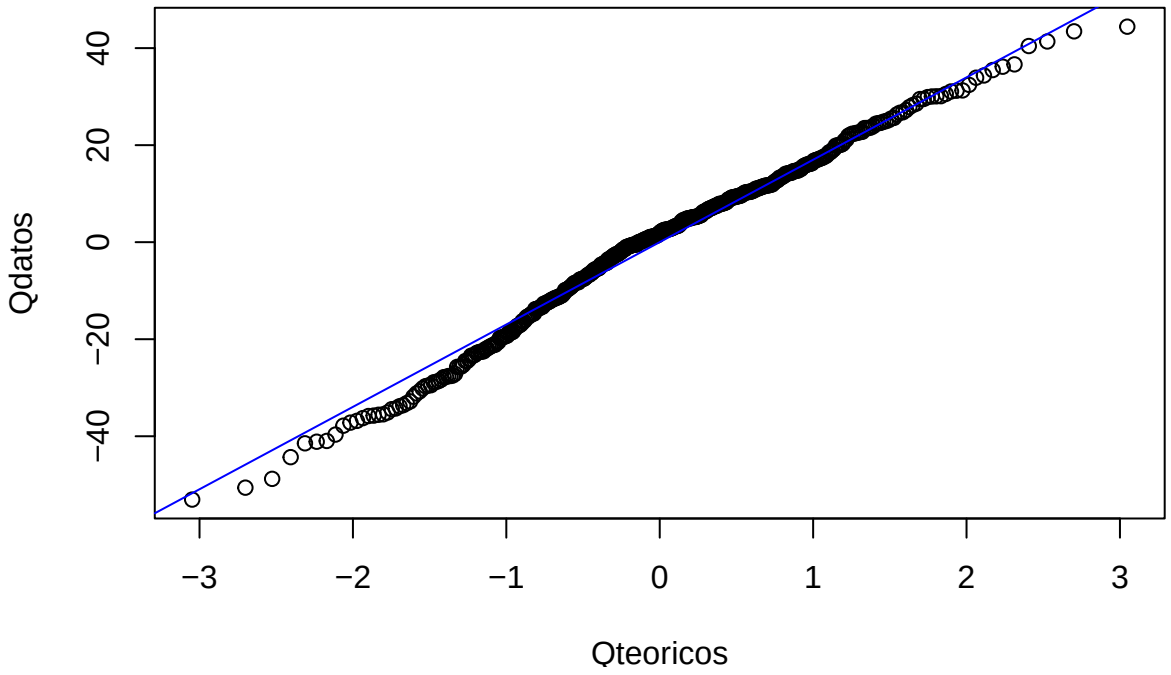
```
plot(Qteoricos,Qdatos)
```



- Según la ayuda de R, *qqline()* pasa por el primer y tercer cuartil.

```
plot(Qteoricos,Qdatos)

# qqline
x1 <- qnorm(0.25)
x2 <- qnorm(0.75)
y1 <- quantile(residuals(m), 0.25)
y2 <- quantile(residuals(m), 0.75)
# mas general
b = (y2-y1)/(x2-x1) # pendiente
a = y1 - b*x1 # y1 = a + b*x1
abline(a,b, col = "blue", lwd = 1) # y = a + b*x
```



Para comprobar la normalidad también se puede utilizar un test de bondad de ajuste. Uno de los más utilizados es el test de normalidad de Shapiro:

- H0: los residuos tienen distribucion normal
- H1: los residuos NO tienen distribución normal

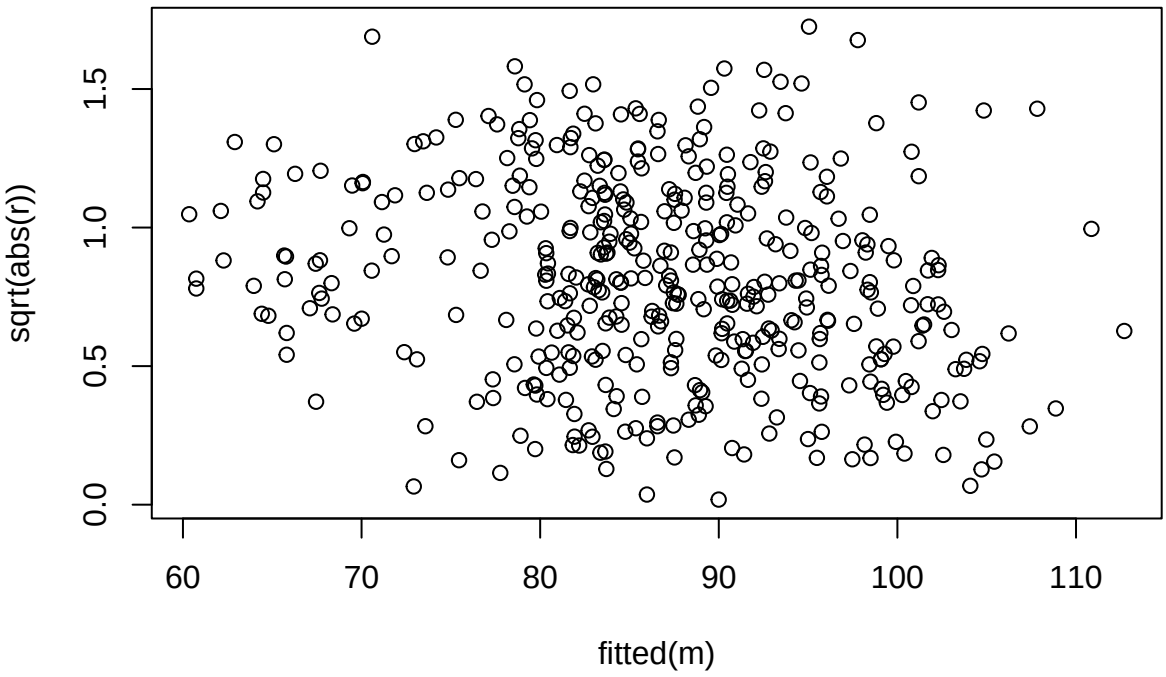
```
shapiro.test(resid(m))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(m)
## W = 0.98949, p-value = 0.003345
```

2.3 Gráfico de residuos estandarizados vs valores predichos.

Se utiliza para comprobar la homocedasticidad:

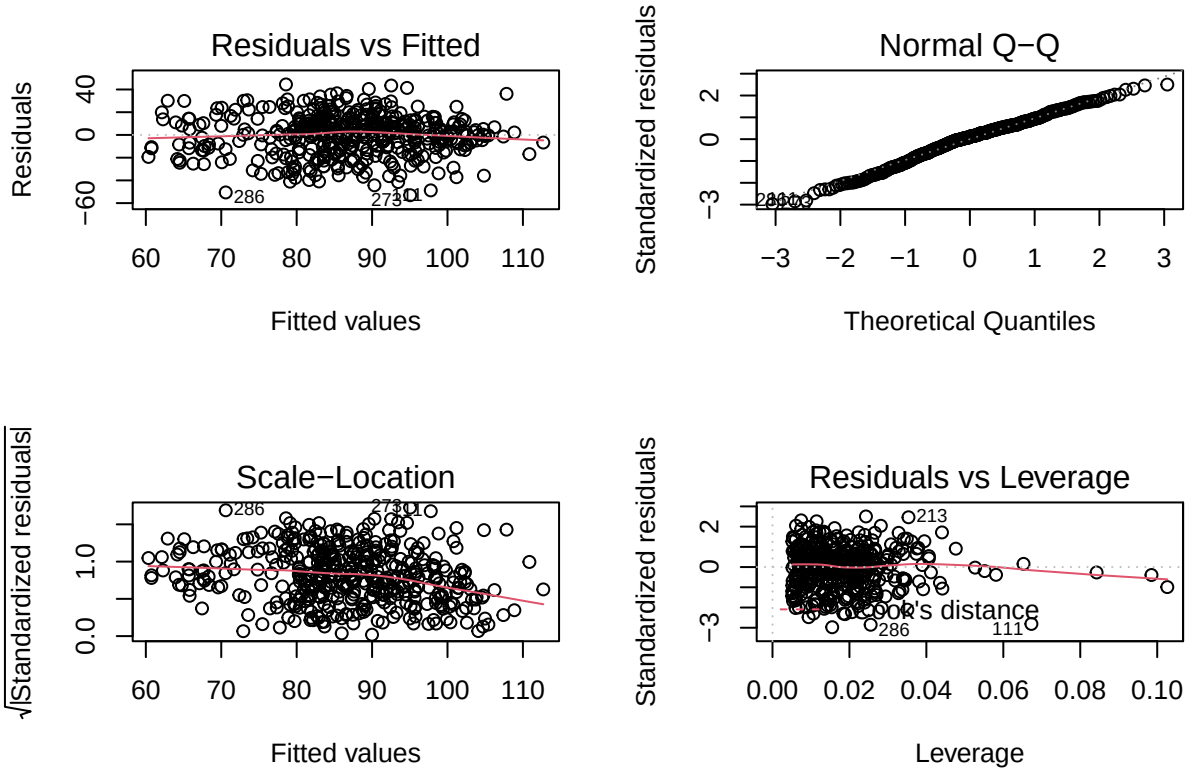
```
sR = sqrt(sum(resid(m)^2)/m$df.residual)
X = model.matrix(m)
H = X %*% solve(t(X) %*% X) %*% t(X)
h = diag(H)
r = resid(m)/(sR*sqrt(1-h))
plot(fitted(m), sqrt(abs(r)))
```



2.4 En R

Los gráficos anteriores se pueden obtener diréctamente en R:

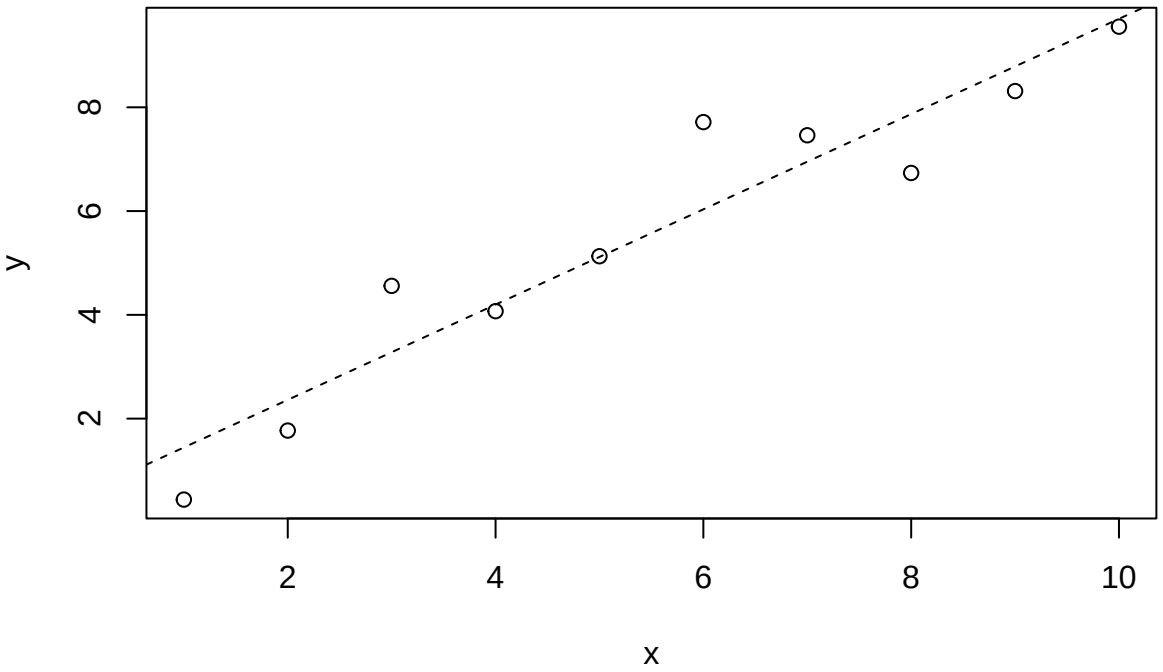
```
par(mfrow=c(2,2))
plot(m)
```



3 Datos atípicos

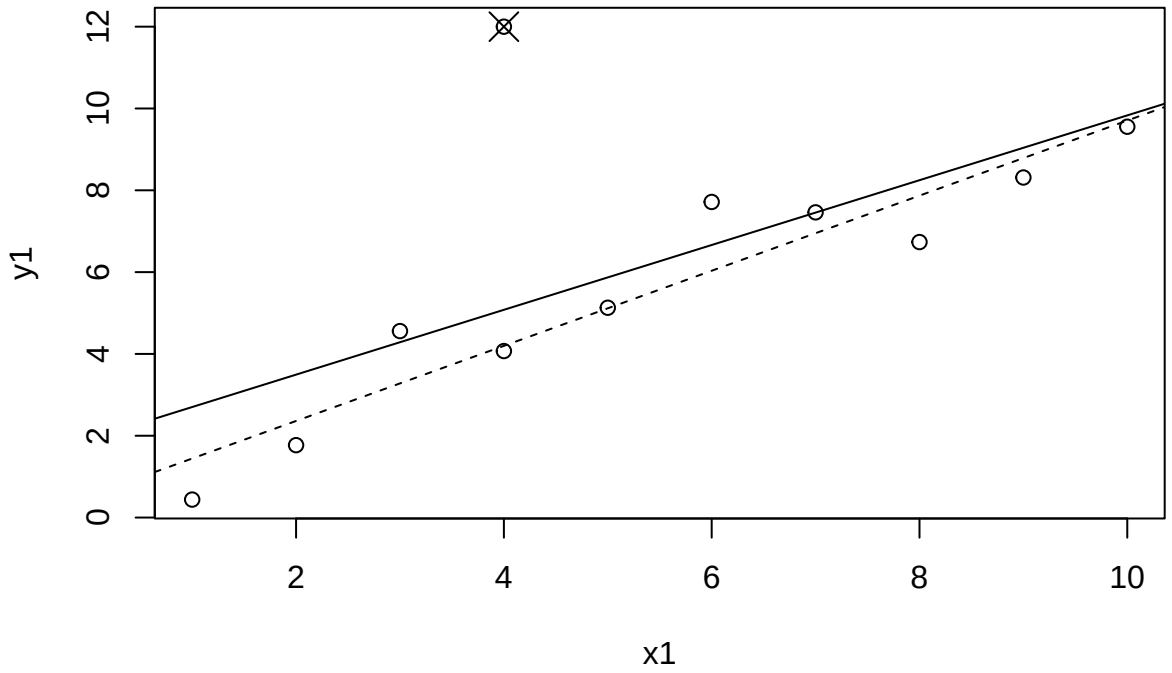
Para trabajar este apartado vamos a utilizar datos simulados:

```
set.seed(123)
x = 1:10
y = x + rnorm(10)
msim = lm(y ~ x)
plot(x,y)
abline(msim, lty = 2)
```



Un atípico es aquel dato que tiene una Y diferente con respecto a resto de los datos.

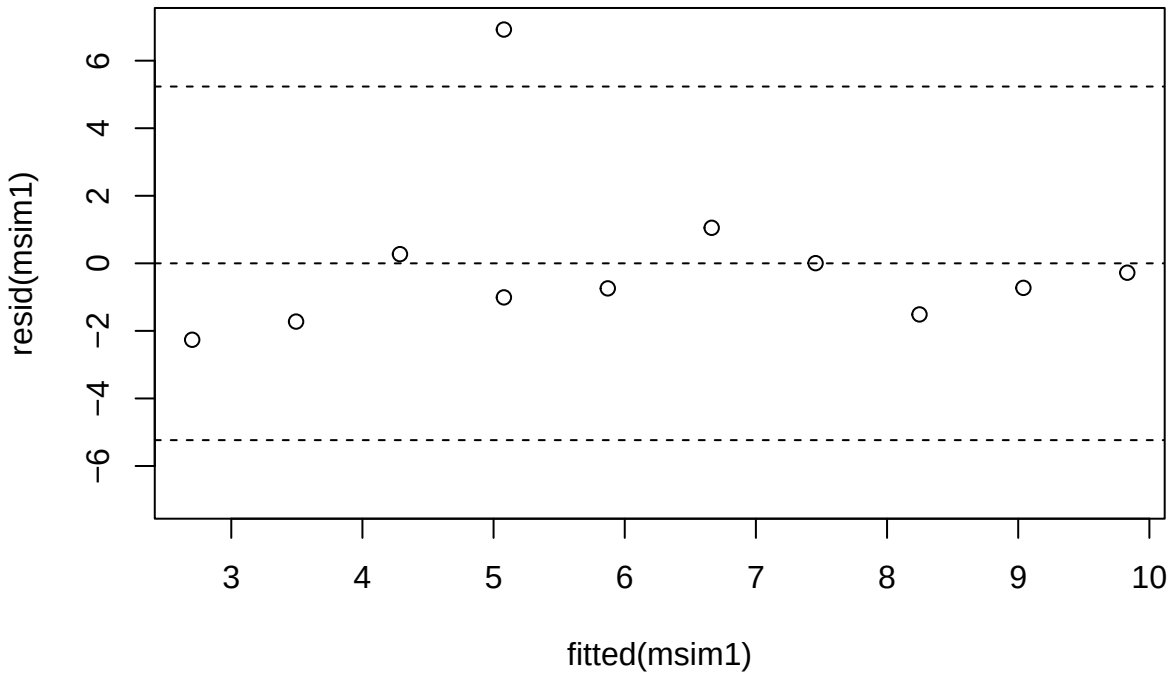
```
x1 = c(x,4)
y1 = c(y,12)
msim1 = lm(y1 ~ x1)
plot(x1,y1)
abline(msim, lty = 2)
abline(msim1)
points(4, 12, pch = 4, cex = 2)
```



Como se observa, los datos atípicos pueden modificar la estimación de los parámetros del modelo. Podemos considerar que el residuo es atípico si

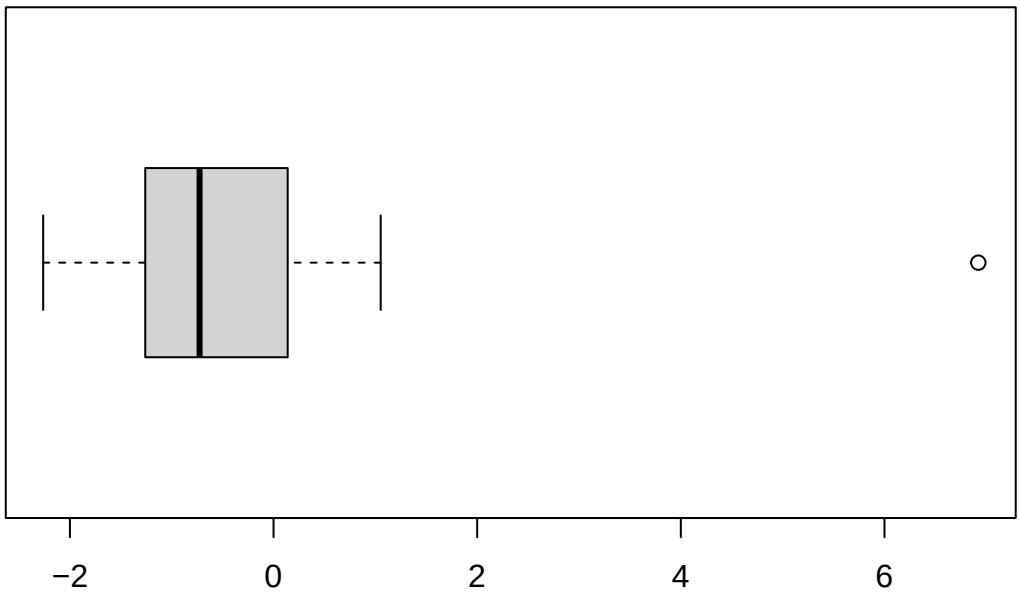
- está fuera de las bandas $\pm 2\hat{s}_R^2$, ya que en una normal estaría el 95% de los datos.

```
plot(fitted(msim1), resid(msim1), ylim = c(-7,7))
sR = sqrt( sum(resid(msim1)^2)/(length(x1)-2) )
abline(h = 0, lty = 2)
abline(h = 2*sR, lty = 2)
abline(h = -2*sR, lty = 2)
```



- También se puede utilizar el criterio del gráfico boxplot:

```
boxplot(resid(msim1), horizontal = T)
```



Este es un punto atípico pero no cambia mucho la estimación de los parámetros del modelo.

Los atípicos pueden deberse a errores en la toma de los datos. En este caso se pueden eliminar del análisis. Si no estamos seguros de que es un error, no se aconseja eliminar los atípicos:

La NASA lanzó el satélite Nimbus 7 para registrar datos atmosféricos. Tiempo después, el British Antarctic Survey observó un descenso importante del ozono atmosférico en la Artártida que no se había detectado en los datos recogidos por dicho satélite. La causa fue que el sistema de recogida de datos no guardaba los datos que eran demasiado bajos ya que los consideraba errores. Esto motivó que la detección del agujero en la capa de ozono se retrasara varios años.