Comprobación de las hipótesis del modelo: diagnosis

Contents

1 Hipótesis del modelo y residuos

Recordamos el modelo:

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

El modelo se basa en $u_i \sim N(0, \sigma^2)$, y pueden describirse como:

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

Los errores no son observables (solo se observan las y_i), 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.

Por otro lado:

$$E[e] = 0$$

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

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. Se definen también los *residuos estandarizados*:

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

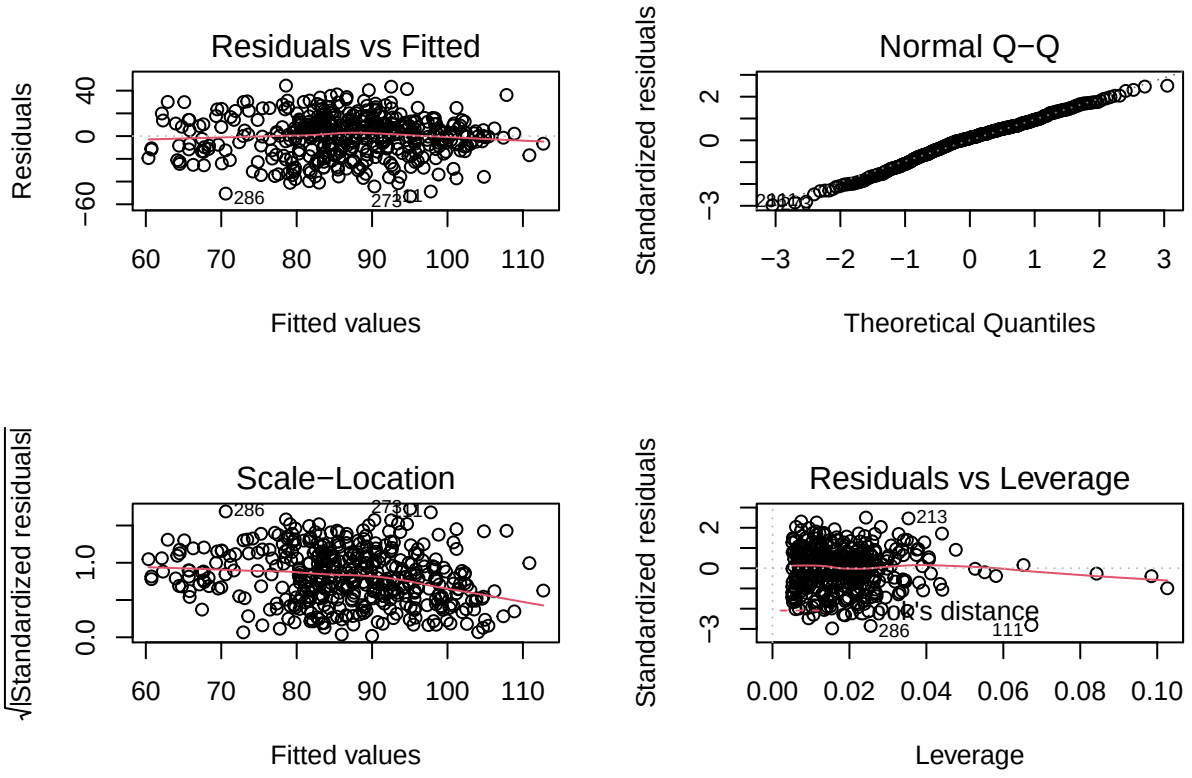
que tienen varianza uno.

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

2 Análisis gráfico de los residuos

2.1 En R

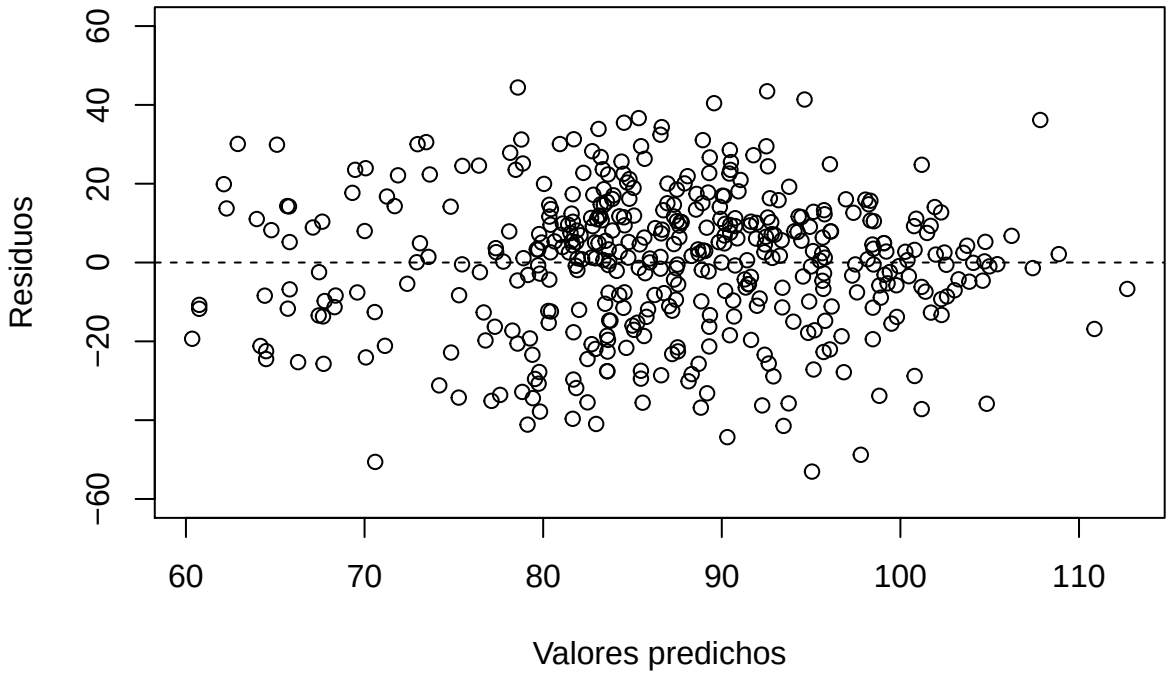
```
d = read.csv("datos/kidiq.csv")
d$mom_hs = factor(d$mom_hs, labels = c("no","yes"))
d$mom_work = factor(d$mom_work)
m = lm(kid_score ~ mom_iq * mom_hs + mom_age + mom_work, data = d)
#
par(mfrow=c(2,2))
plot(m)
```



2.2 Residuos frente a valores predichos

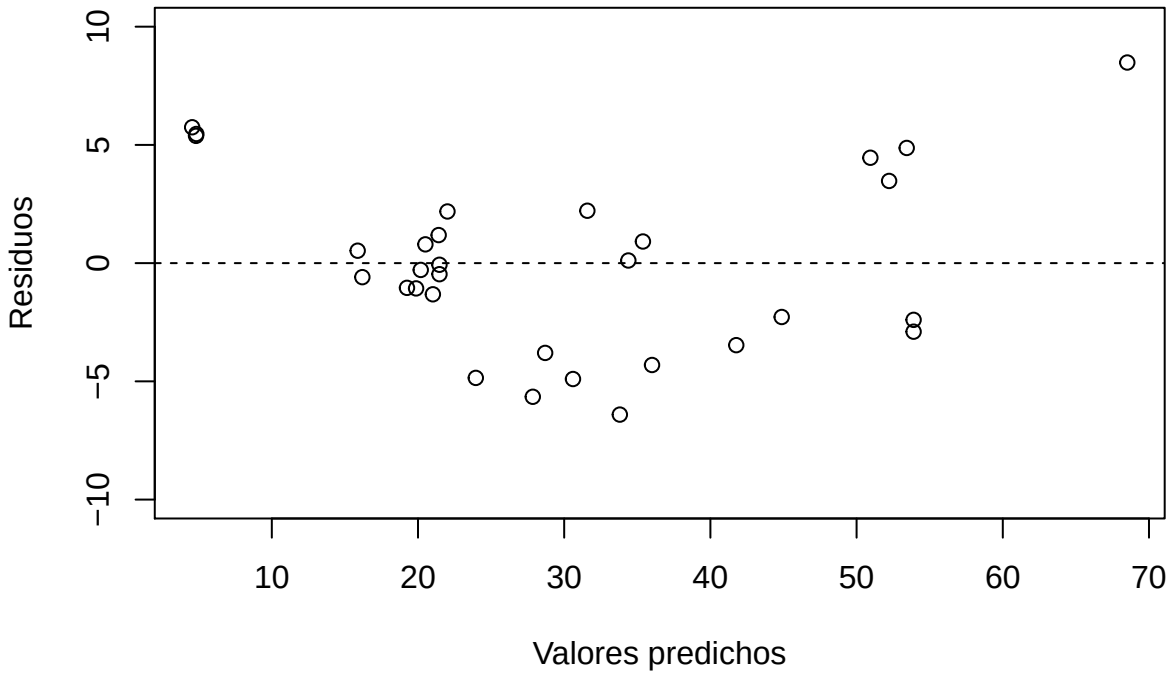
La herramienta más usada es el gráfico de residuos frente a valores predichos

```r
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.
- El caso más frecuente cuando hay heterocedasticidad es que la dispersión de los residuos aumente con el valor de $\hat{y}_i$.
- El caso más frecuente cuando no hay linealidad es que se observe curvatura en los residuos.
- Por ejemplo, sean los datos:

```r
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)
```

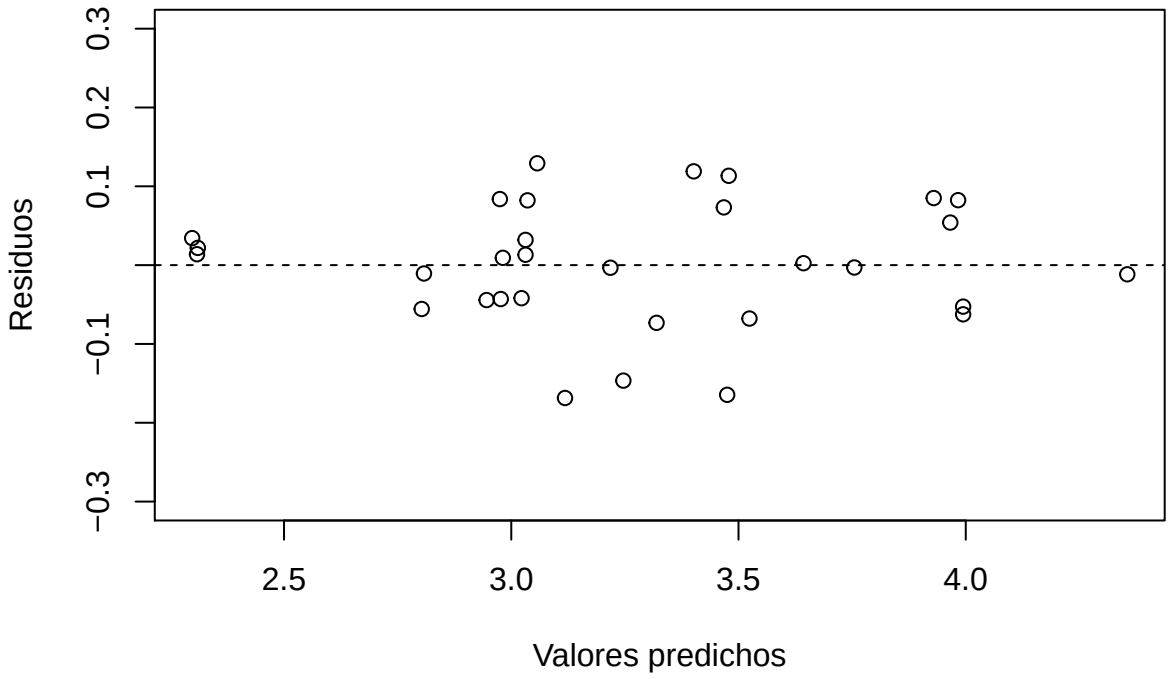


En este caso, no se cumple homocedasticidad ni linealidad.

- 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 simultanenamente.
- 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)
```



2.3 Gráfico Cuantil-Cuantil

Se utiliza para detectar atípicos y verificar la hipótesis de normalidad. En este gráfico se comparan los cuantiles de los datos con los cuantiles teóricos

- Cuantiles de los datos: son los datos ordenados, de menor a mayor.

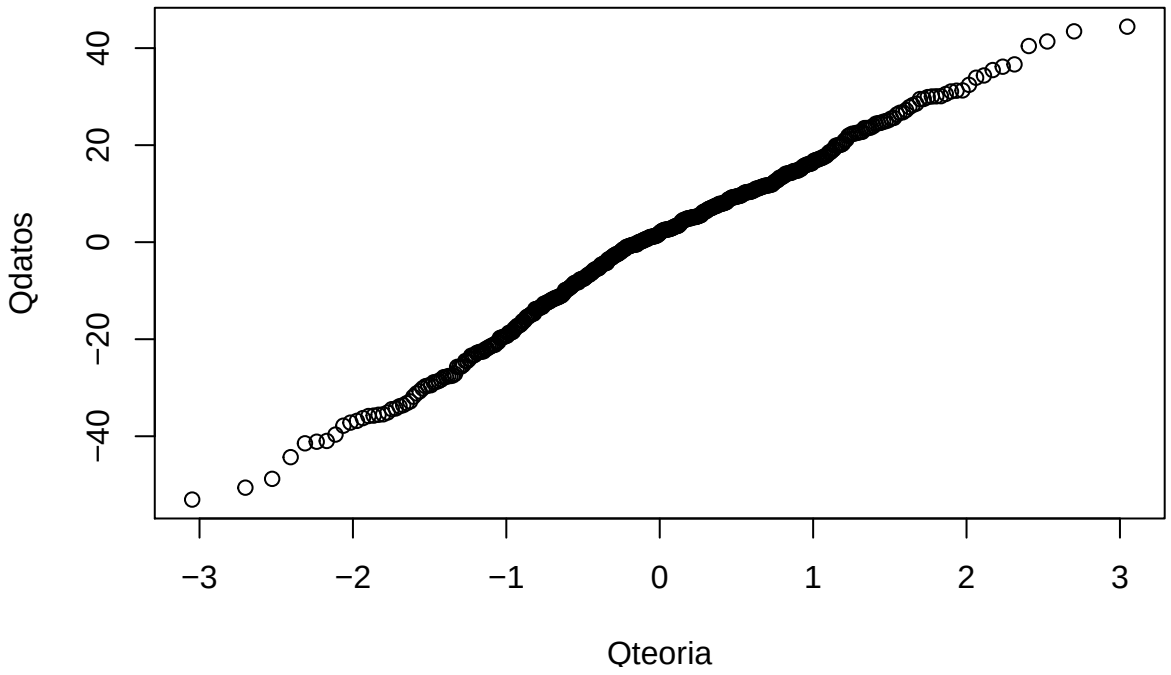
```
Qdatos = sort(residuals(m))
```

- Cuantiles teóricos: son los cuantiles de la N(0,1) correspondientes a $(i-0.5)/n$, donde n es el número de datos, $i = 1, 2, \cdots, n$.

```
n = length(residuals(m))
i = 1:n
q = (i-0.5)/n
Qteoria = qnorm(q)
```

- Se representa Qteoria *vs* Qdatos.

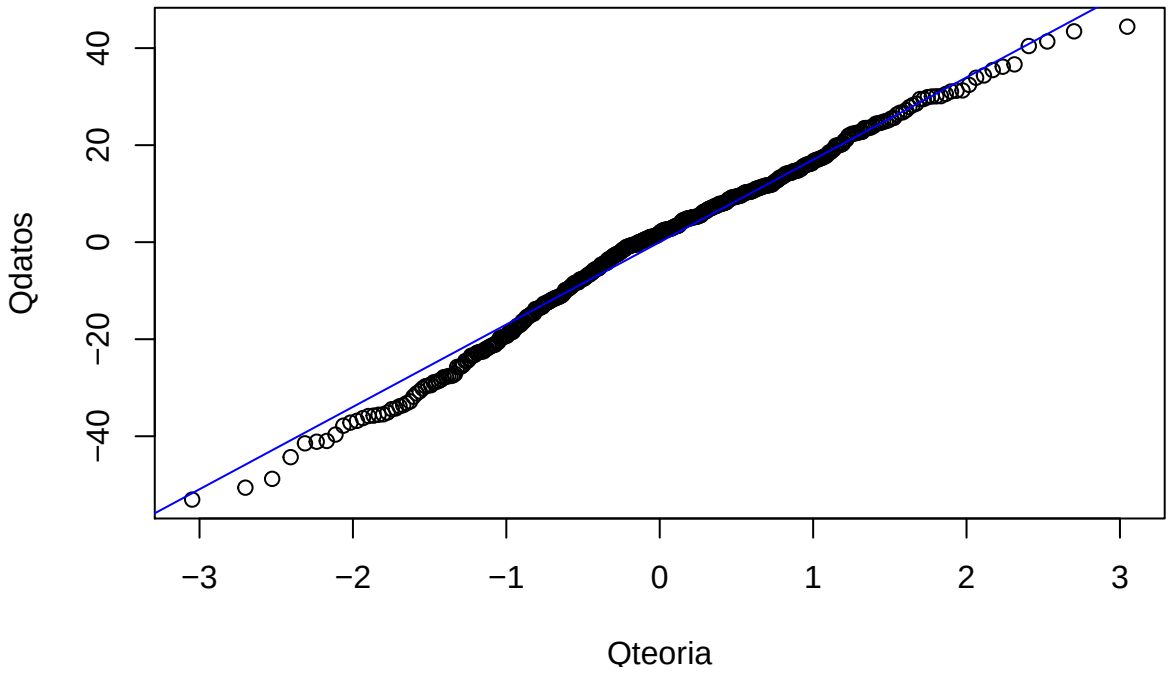
```
plot(Qteoria,Qdatos)
```



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

```
plot(Qteoria,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

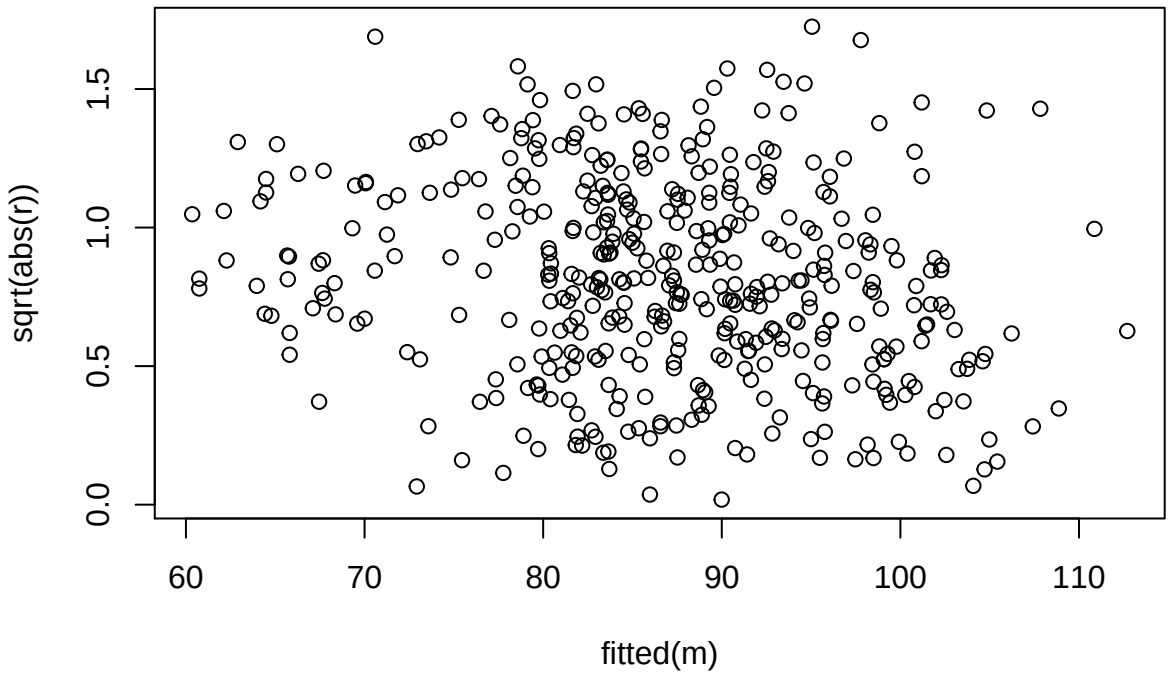
```
shapiro.test(resid(m))
```

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

2.4 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)))
```

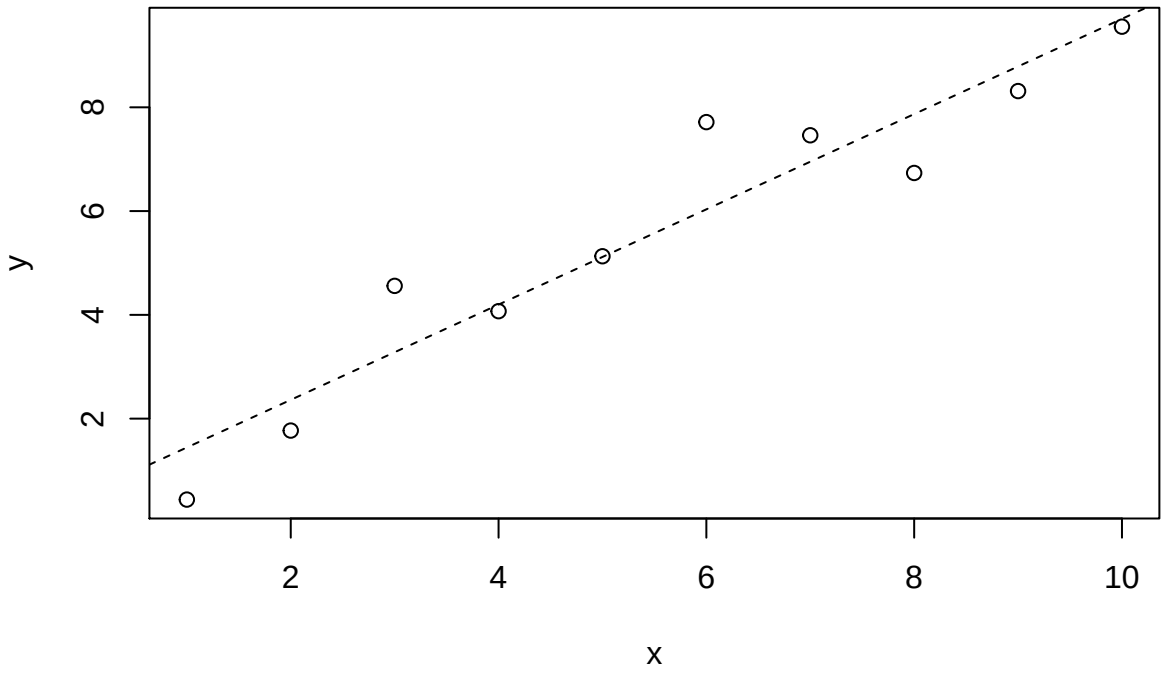


2.5 Gráfico de leverage vs residuos estandarizados

2.5.1 Datos simulados

Para trabajar este apartado vamos a utilizar unos datos simulados:

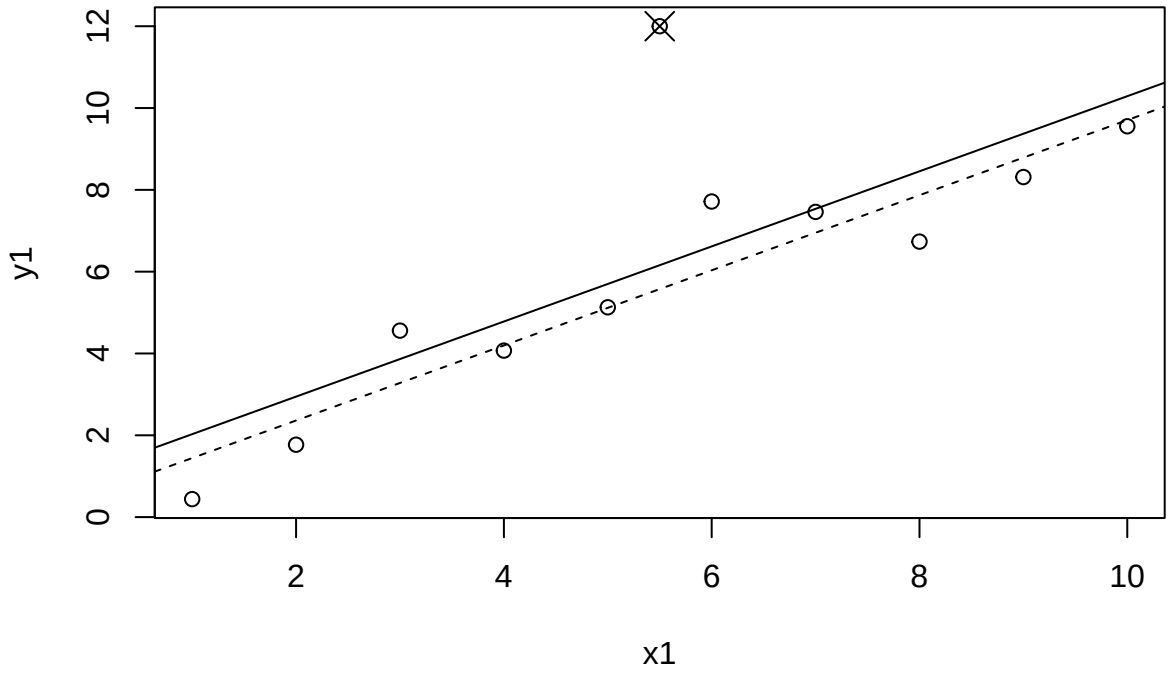
```
set.seed(123)
x = 1:10
y = x + rnorm(10)
msim = lm(y ~ x)
plot(x,y)
abline(msim, lty = 2)
```



2.5.2 Atípicos

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

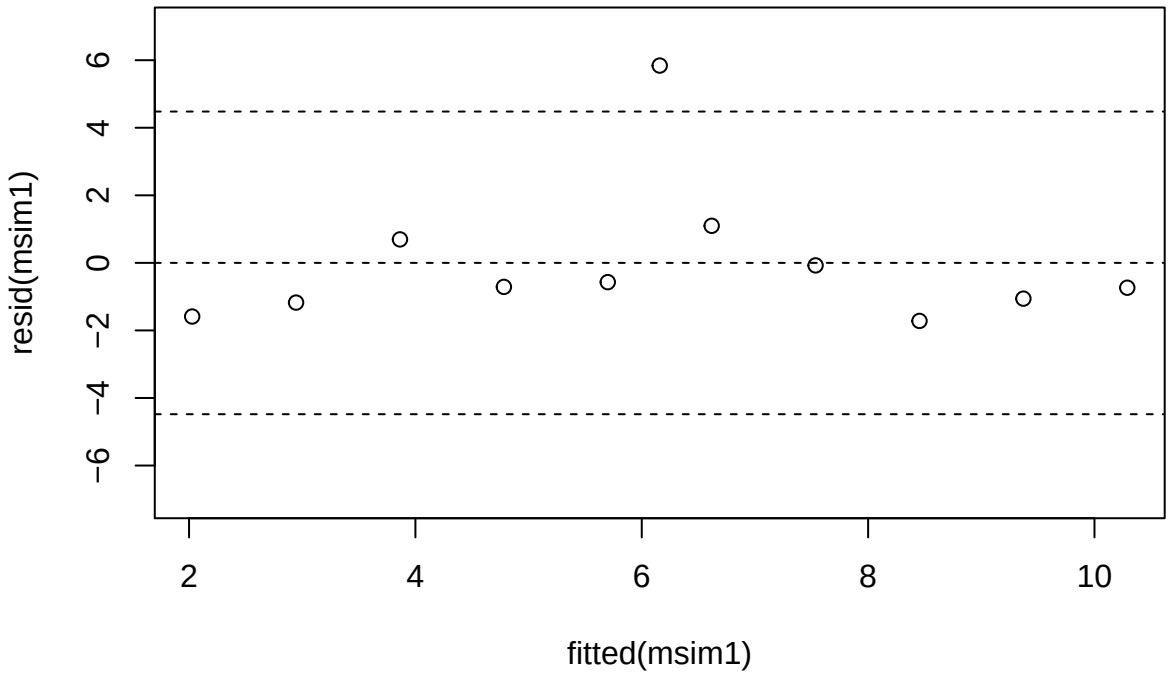
```
x1 = c(x,5.5)
y1 = c(y,12)
msim1 = lm(y1 ~ x1)
plot(x1,y1)
abline(msim, lty = 2)
abline(msim1)
points(5.5, 12, pch = 4, cex = 2)
```



Podemos considerar que el residuo es atípico si

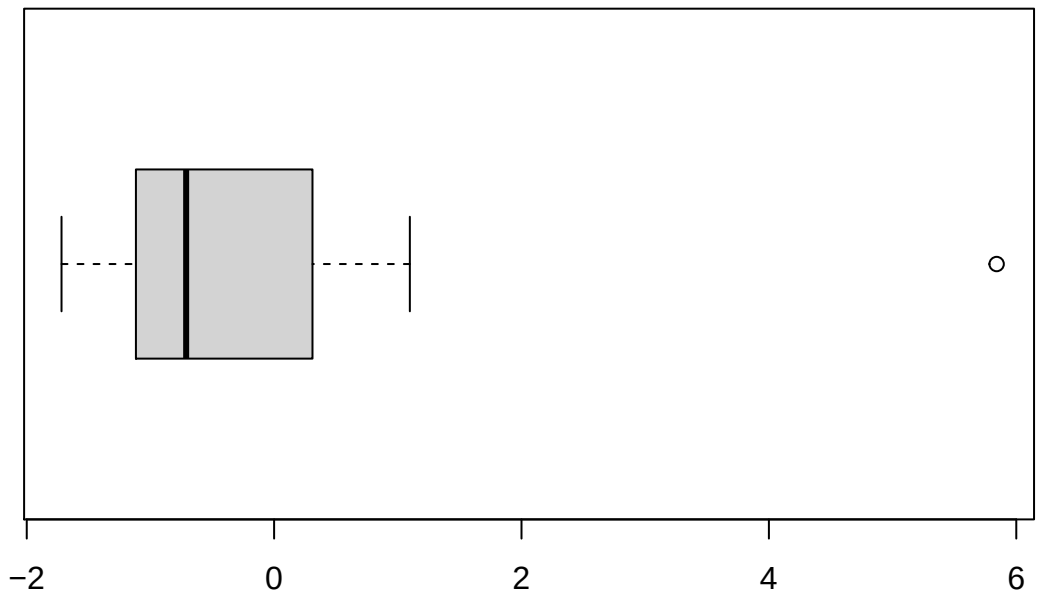
- está fuera de las bandas $\pm 2\hat{s}_R^2$, ya que en una normal estaría el 95% de los datos.

```
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.

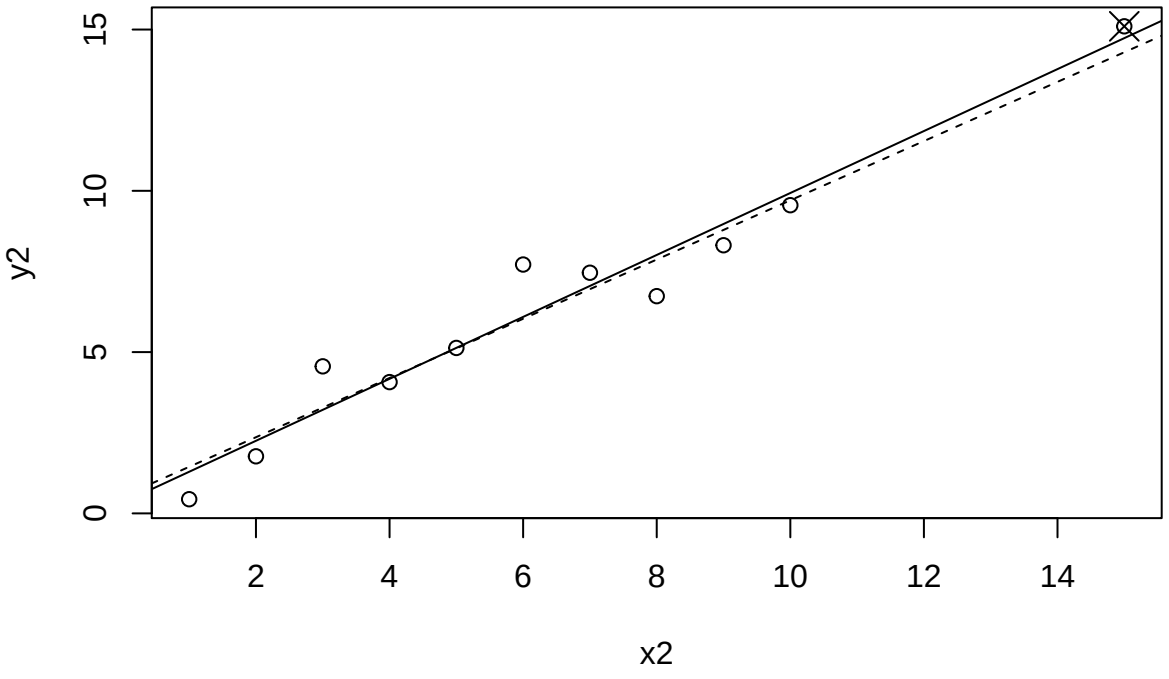
Los atípicos pueden deberse a errores en la toma de los datos. En este caso se pueden eliminar del análisis. Si no estamos seguros de que es un error, no se aconseja eliminar los atípicos:

La NASA lanzó el satélite Nimbus 7 para toma de datos atmosféricos. Tiempo después, el British Antarctic Survey observó un descenso importante del ozono atmosférico en la Artártida que no se registró en los datos recogidos por el satélite. La causa fue que el sistema de recogido 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 retrasó varios años

2.5.3 Leverage

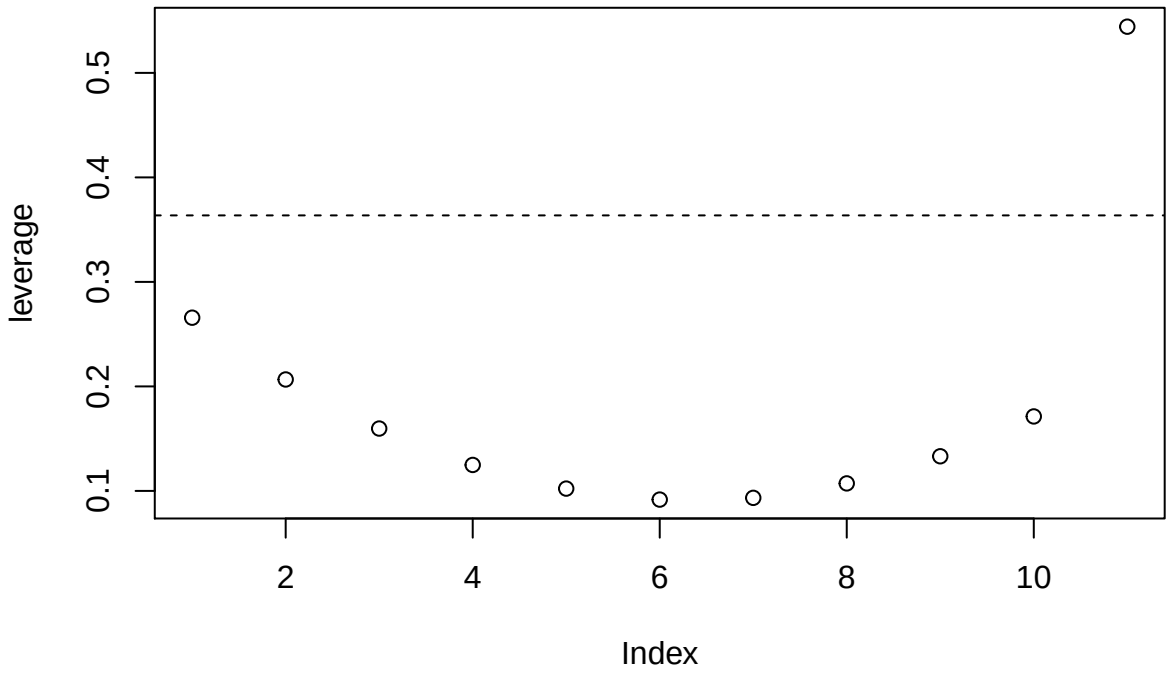
- Un dato leverage es aquel que tiene un X elevado con respecto al resto de los datos.
- Una manera de medir el leverage es mediante h_{ii}.
- Tradicionalmente se ha considerado que los puntos con leverage mayor que 2*p/n hay que mirarlos con cuidado, donde $p = \sum h_{ii}$.

```
x2 = c(x,15)
y2 = c(y, 15.1)
msim2 = lm(y2 ~ x2)
plot(x2,y2)
abline(msim, lty = 2)
abline(msim2)
points(15, 15.1, pch = 4, cex = 2)
```



Calculamos el leverage de los datos:

```
X = model.matrix(msim2)
H = X %*% solve(t(X) %*% X) %*% t(X)
h = diag(H)
n = length(x2)
p = sum(h)
plot(h, ylab = "leverage")
abline(h = 2*p/n, lty = 2)
```

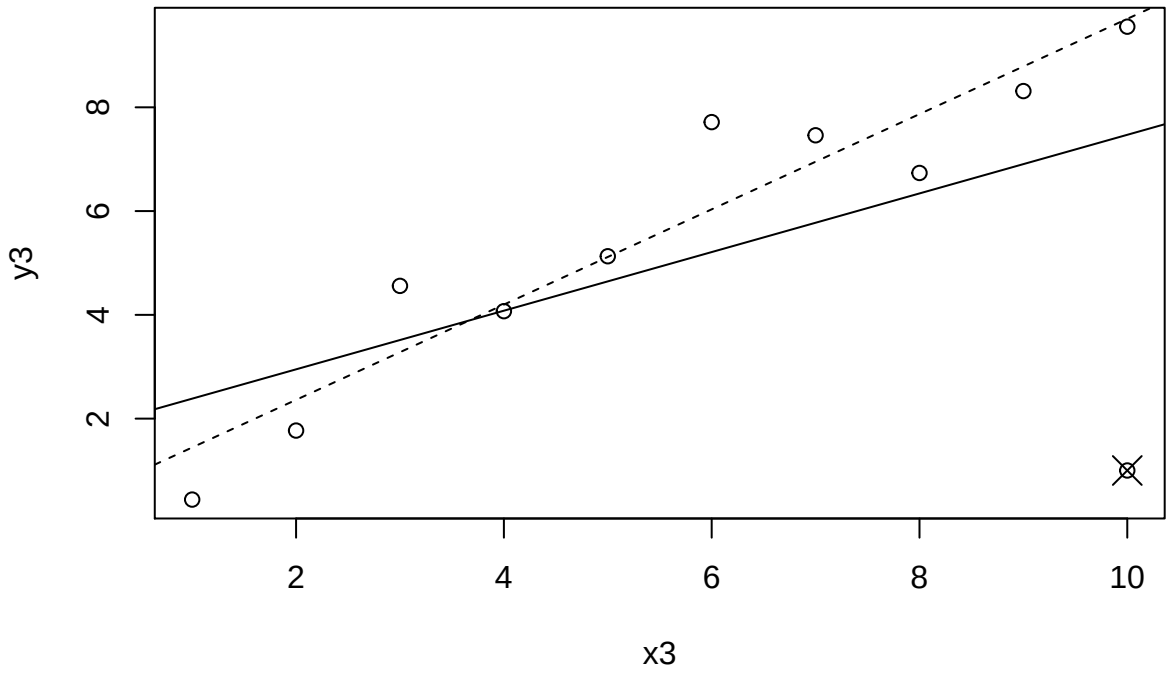


Este punto tiene alto leverage pero no es atípico ni influyente. En principio estos puntos son potencialmente problemáticos, pero no tienen por qué.

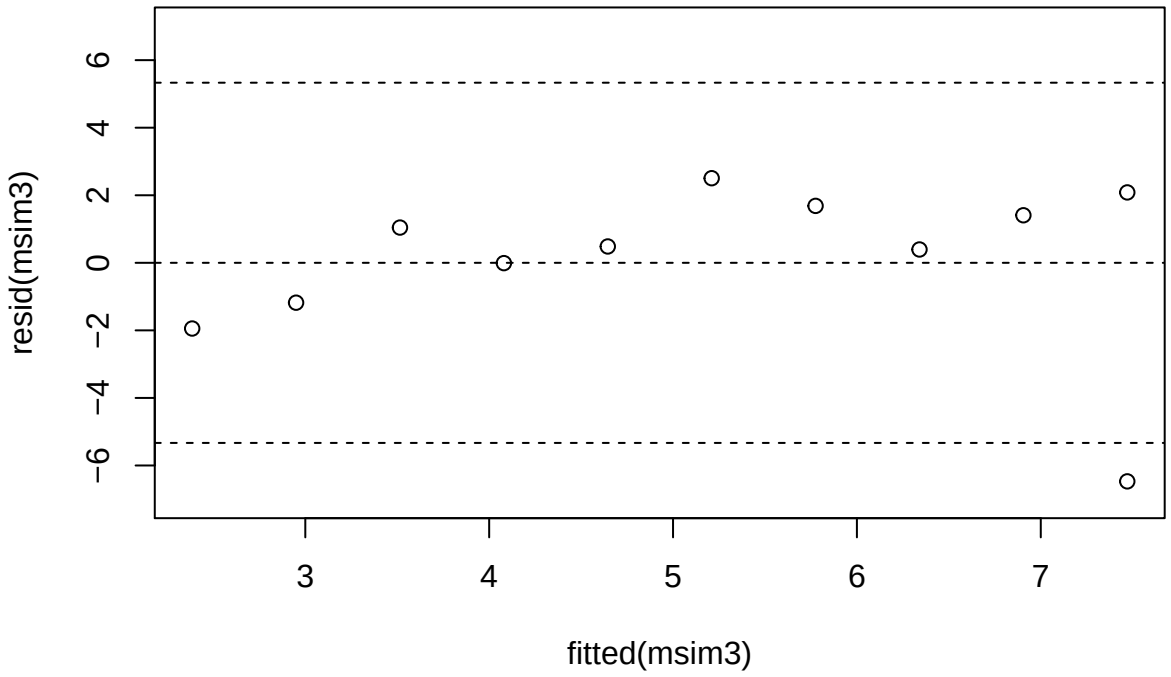
2.5.4 Puntos influyentes

Son aquellos que al quitarlos del modelo cambia mucho la estimación de los parámetros:

```
x3 = c(x,10)
y3 = c(y, 1)
msim3 = lm(y3 ~ x3)
plot(x3,y3)
abline(msim, lty = 2)
abline(msim3)
points(10, 1, pch = 4, cex = 2)
```




```
plot(fitted(msim3), resid(msim3), ylim = c(-7,7))
sR = sqrt( sum(resid(msim3)^2)/(length(x3)-2) )
abline(h = 0, lty = 2)
abline(h = 2*sR, lty = 2)
abline(h = -2*sR, lty = 2)
```



Este punto es influyente y atípico al mismo tiempo.

Hay varias formas de medir la influencia de los datos. El método más popular es el estadístico de Cook:

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

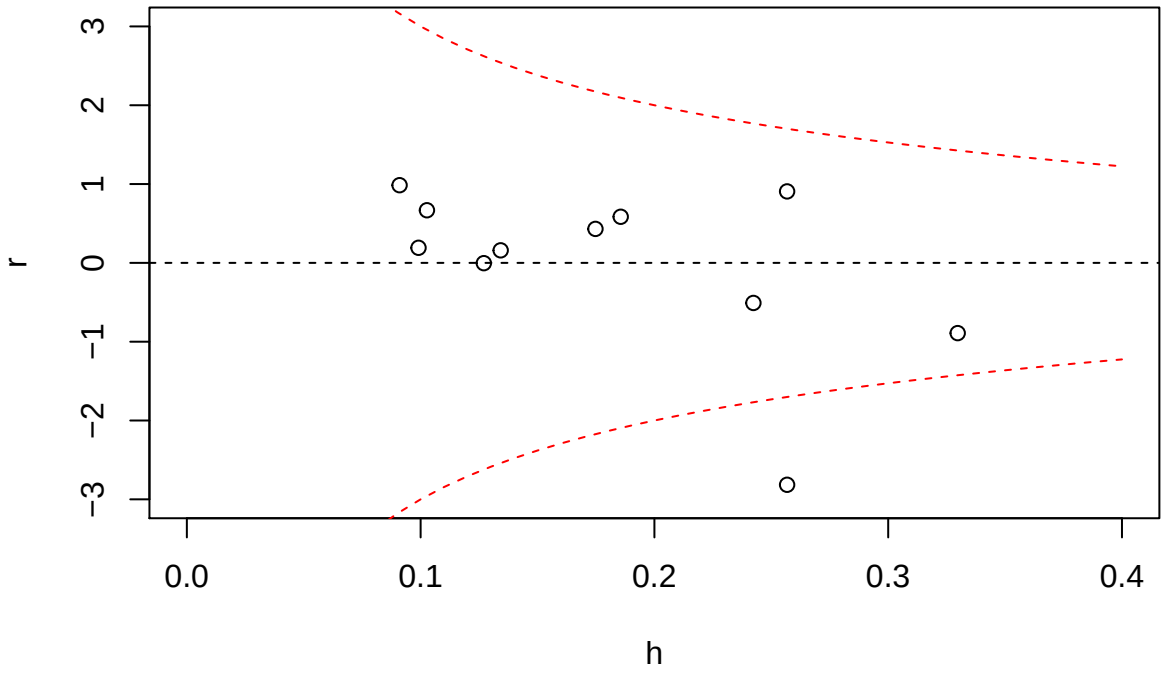
Por lo tanto, para considerar un punto como influyente D_i tiene en cuenta el residuo y el leverage.

De la ecuación de Cook tenemos:

$$r_i = \pm\sqrt{\frac{pD_i(1-h_{ii})}{h_{ii}}}$$

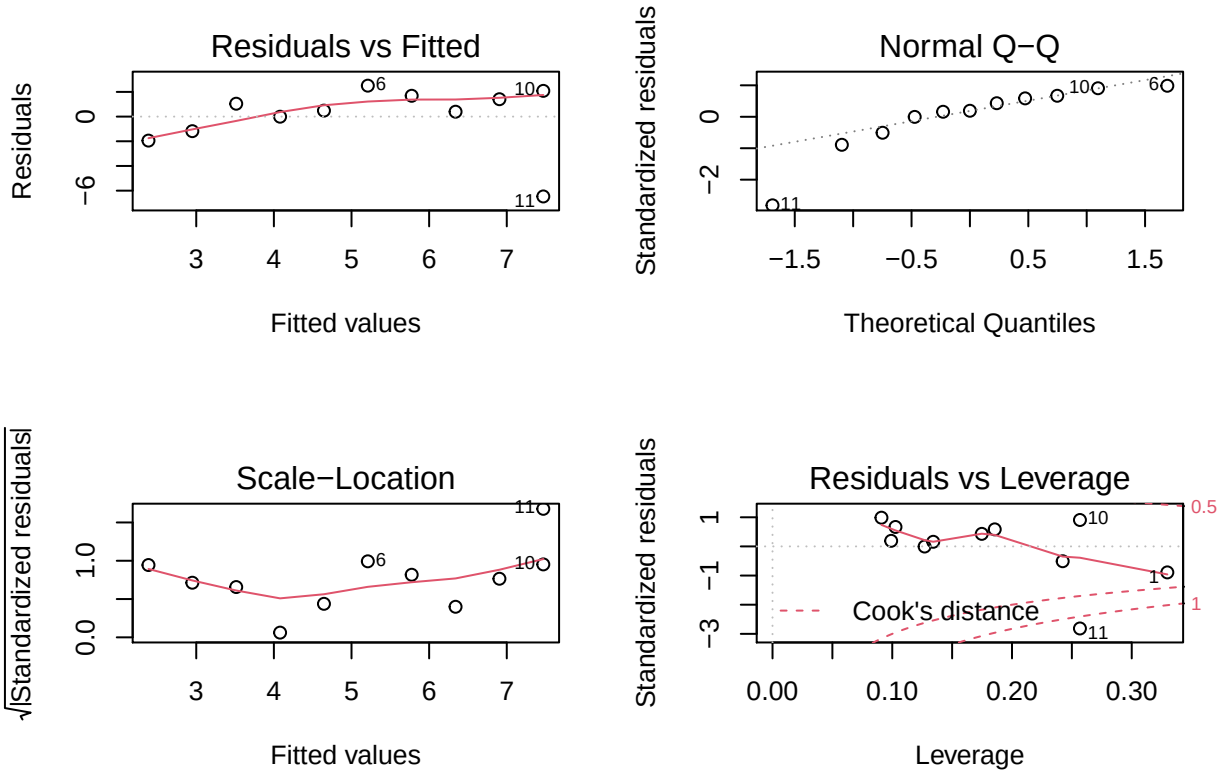
Dando valores a Di:

```
X = model.matrix(msim3)
H = X %*% solve(t(X) %*% X) %*% t(X)
h = diag(H)
n = length(x3)
p = sum(h)
r = resid(msim3)/(sR*sqrt(1-h))
plot(h,r, xlim = c(0,0.4), ylim = c(-3,3))
abline(h = 0, lty = 2)
#
ha = seq(0, 0.4, 0.01)
Da = 0.5
ra = sqrt(p*Da*(1-ha)/ha)
lines(ha,ra, lty = 2, col = "red")
lines(ha,-ra, lty = 2, col = "red")
```



En R:

```
par(mfrow = c(2,2))
plot(msim3)
```



Los puntos que están fuera de estos límites son influyentes.