Ejemplo de regresión lineal

Contents

1 Datos

```r
d = read.csv("datos/Advertising.csv")
str(d)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

- sales: en miles de unidades
- TV, radio, newspaper: presupuesto de publicidad, en miles de dolares

2 Modelo 1

2.1 Estimación

```r
m1 = lm(sales ~ TV + radio + newspaper, data = d)
summary(m1)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
```

```
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_0 = 2.94$. Como p-valor $< \alpha$, es un parámetro significativo.
- $\hat{\beta}_1 = 0.05$. Si mantenemos la inversión en *radio* y *newspaper* constante, un incremento de 1000 $ en TV, por término medio supone un aumento en las ventas de 0.05*1000 = 50 unidades. Según el pvalor, es un parámetro significativo.
- $\hat{\beta}_2 = 0.19$. Si mantenemos la inversión en TV y *newspaper* constante, un incremento de 1000 $ en TV, por término medio supone un aumento en las ventas de 190 unidades. Según el pvalor, es un parámetro significativo.
- $\hat{\beta}_3$ = -0.001. Según el pvalor, es un parámetro **NO** significativo, luego no influye en las ventas. Sin embargo, si analizamos la regresión simple de *newspaper*:

La variable *newspaper* ha resultado no significativa. Puede deberse a dos cosas:

- Que no está relacionada con *sales*. Esto lo comprobamos con una regresión simple:

```r
m1a = lm(sales ~ newspaper, data = d)
summary(m1a)
```

```
##
## Call:
## lm(formula = sales ~ newspaper, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
## newspaper    0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

luego si está relacionada.

- que no aporte información adicional a la que ya aportan TV y *radio*. Esto sucede si la correlación con alguna de las variables es considerable:

```r
cor(d[,-1])
```

```
##                   TV      radio  newspaper     sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## radio     0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Como vemos, la correlación entre radio y newspaper es 0.35, lo que indica que en los mercados donde se invierte en *radio* también se invierte en *newspaper*. Luego parte de la información contenida en *newspaper* ya está contenida en *radio*, luego no aporta información significativa si inclimos las dos variables en el mismo modelo:

```r
m1b = lm(sales ~ radio + newspaper, data = d)
summary(m1b)
```

```
##
## Call:
## lm(formula = sales ~ radio + newspaper, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5289  -2.1449   0.7315   2.7657   7.9751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.188920   0.627672  14.640   <2e-16 ***
## radio       0.199045   0.021870   9.101   <2e-16 ***
## newspaper   0.006644   0.014909   0.446    0.656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.284 on 197 degrees of freedom
## Multiple R-squared:  0.3327, Adjusted R-squared:  0.3259
## F-statistic: 49.11 on 2 and 197 DF,  p-value: < 2.2e-16
```

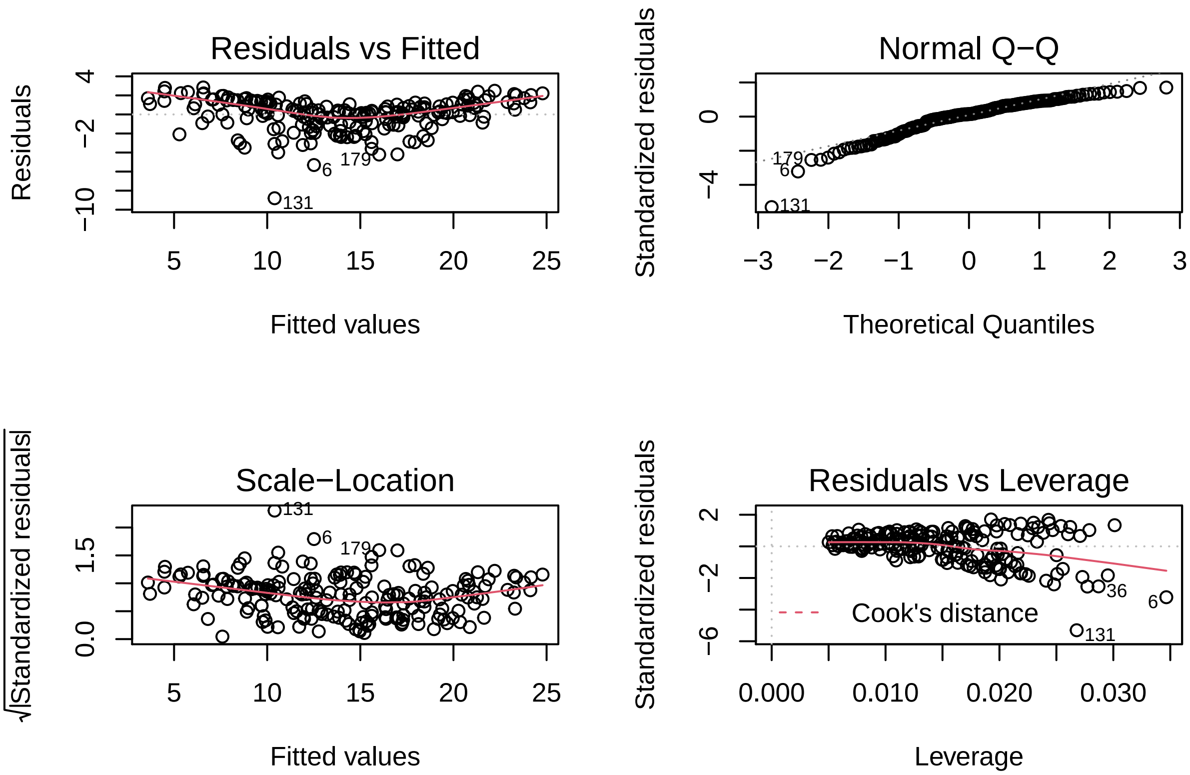
Por tanto, vamos a quitar newspaper:

```r
m2 = lm(sales ~ TV + radio, data = d)
summary(m2)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

2.2 Comprobación de las hipótesis del modelo

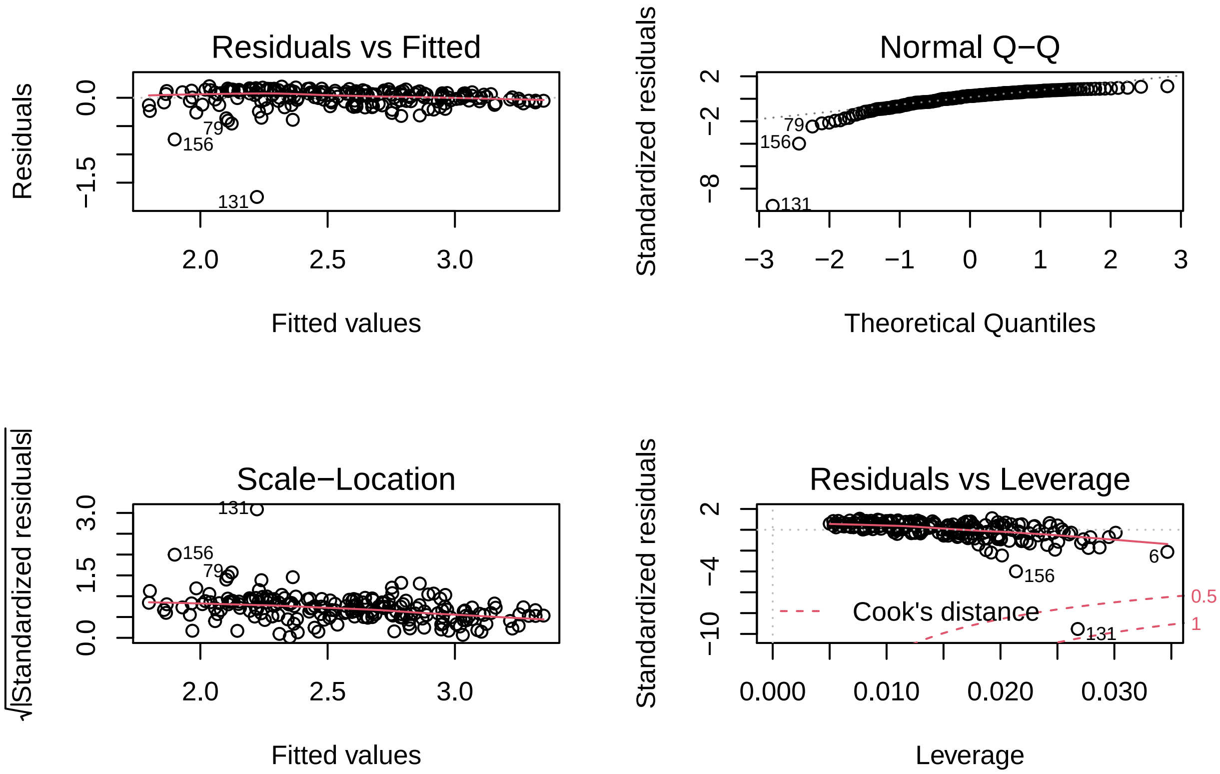
```r
par(mfrow=c(2,2))
plot(m2)
```



Como vemos, no se cumple linealidad. Una solución simple consiste en usar transformaciones no-lineales de las X. Las más comunes son: $\log(X), \sqrt{X}, X^2, 1/X$.

Podemos comprobar que ninguna de ellas corrige la linealidad. Sin embargo, si aplicamos logaritmos a la variable respuesta:

```r
m3 = lm(log(sales) ~ TV + radio, data = d)
par(mfrow=c(2,2))
plot(m3)
```




```r
summary(m3)
```

```
##
## Call:
## lm(formula = log(sales) ~ TV + radio, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75225 -0.05628  0.04626  0.10554  0.20542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.7450782  0.0326640   53.42   <2e-16 ***
## TV          0.0036731  0.0001542   23.82   <2e-16 ***
## radio       0.0119849  0.0008918   13.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1865 on 197 degrees of freedom
## Multiple R-squared:  0.7995, Adjusted R-squared:  0.7974
## F-statistic: 392.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

se cumplen razonablemente las hipótesis del modelo. Podemos responder a las preguntas:

- ¿Hay relación entre el gasto en publicidad y las ventas?

Podemos utilizar el contraste general de regresión $H_0 : \beta_1 = \beta_2 = 0$, con pvalor < 2.2e-16., luego hay evidencia clara de la relación entre gasto y ventas.

- ¿Es grande esa relación?

Podemos mirar el $R^2 = 0.80$, luego estamos explicando el 80% de la variabilidad de los datos con este modelo.

- ¿Que medios contribuyen a las ventas?

Viendo los contrastes individuales, contribuyen la radio y la TV, pero no newspaper.

- ¿Como de grande es el efecto de cada medio?

Mirando las β_i, tiene 3 veces más efecto invertir en radio que en TV.

- ¿Cual es la precisión de estos valores?

Podemos mirar los *se*:

```
sqrt(diag(vcov(m3)))
```

```
##  (Intercept)           TV        radio
## 0.0326640167 0.0001542146 0.0008917725
```

O los intervalos de confianza:

```
confint(m3)
```

```
##                   2.5 %      97.5 %
## (Intercept) 1.680662146 1.809494191
## TV          0.003368933 0.003977179
## radio       0.010226276 0.013743568
```

- ¿Como de precisas son las predicciones?

Valor medio predicho - intervalo de confianza del valor medio predicho:

```
xp = data.frame(TV = 50, radio = 40, newspaper = 60)
exp(predict(m3, newdata = xp, level = 0.95, interval="confidence"))
```

```
##        fit      lwr      upr
## 1 11.11314 10.57021 11.68395
```

Valor medio predicho - intervalo de predicción:

```
exp(predict(m3, newdata = xp, level = 0.95, interval="prediction"))
```

```
##        fit      lwr      upr
## 1 11.11314 7.667231 16.10774
```

3 Extensiones del modelo lineal

El problema con los residuos también se puede deber a que no se han incluido los regresores adecuados:

3.1 Términos de interacción

```
m4 = lm(sales ~ TV * radio, data = d)
summary(m4)
```

```
## 
## Call:
## lm(formula = sales ~ TV * radio, data = d)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.3366 -0.4028  0.1831  0.5948  1.5246 
## 
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

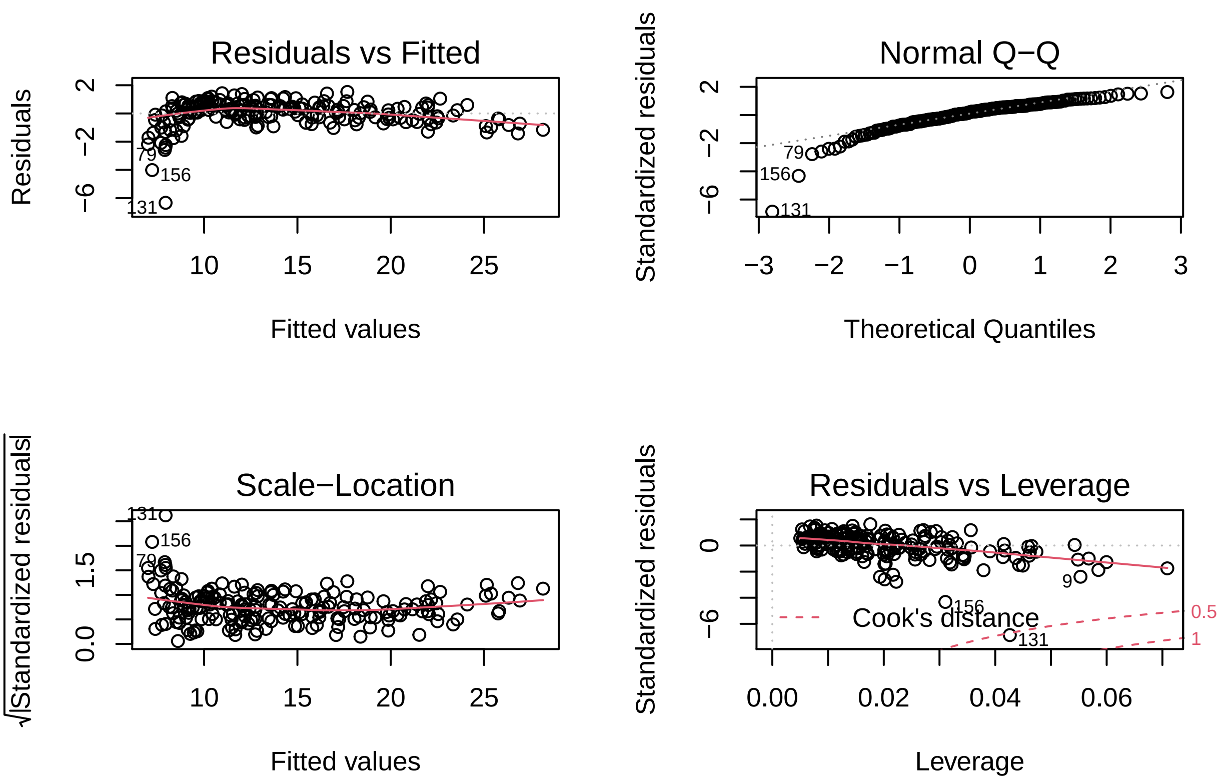
El modelo mejora considerablemente el R^2. Luego invertir dinero en *radio* también mejora la inversión en TV.

$sales = \beta_0 + \beta_1 * TV + \beta_2 * radio + \beta_3 * TV * radio + u$

$sales = \beta_0 + (\beta_1 + \beta_3 * radio) * TV + \beta_2 * radio + u$

La linealidad también ha mejorado:

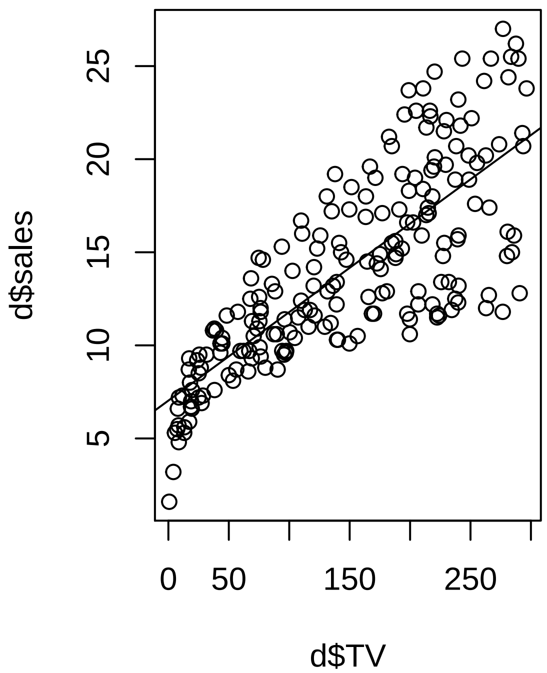
```
par(mfrow=c(2,2))
plot(m4)
```

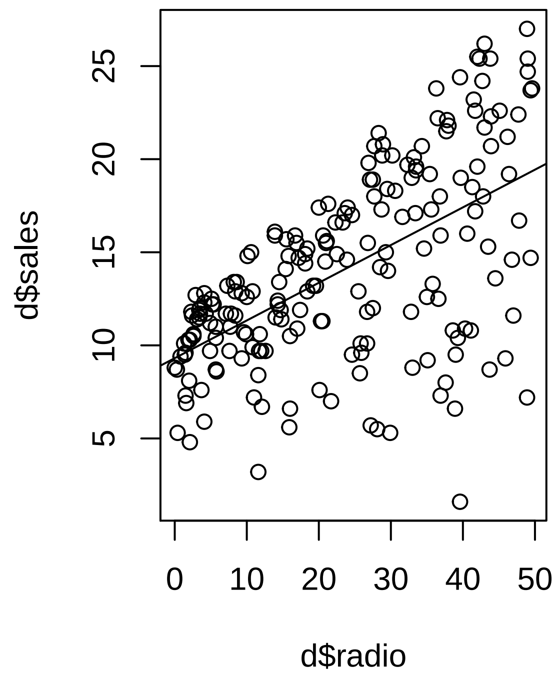


Aún hay que mejorar los residuos.

3.2 Términos no lineales

```
par(mfrow=c(1,2))
plot(d$TV, d$sales)
abline(lm(sales ~ TV, data = d))
plot(d$radio, d$sales)
abline(lm(sales ~ radio, data = d))
```





Parece que la relación con TV es de orden 2 o 3:

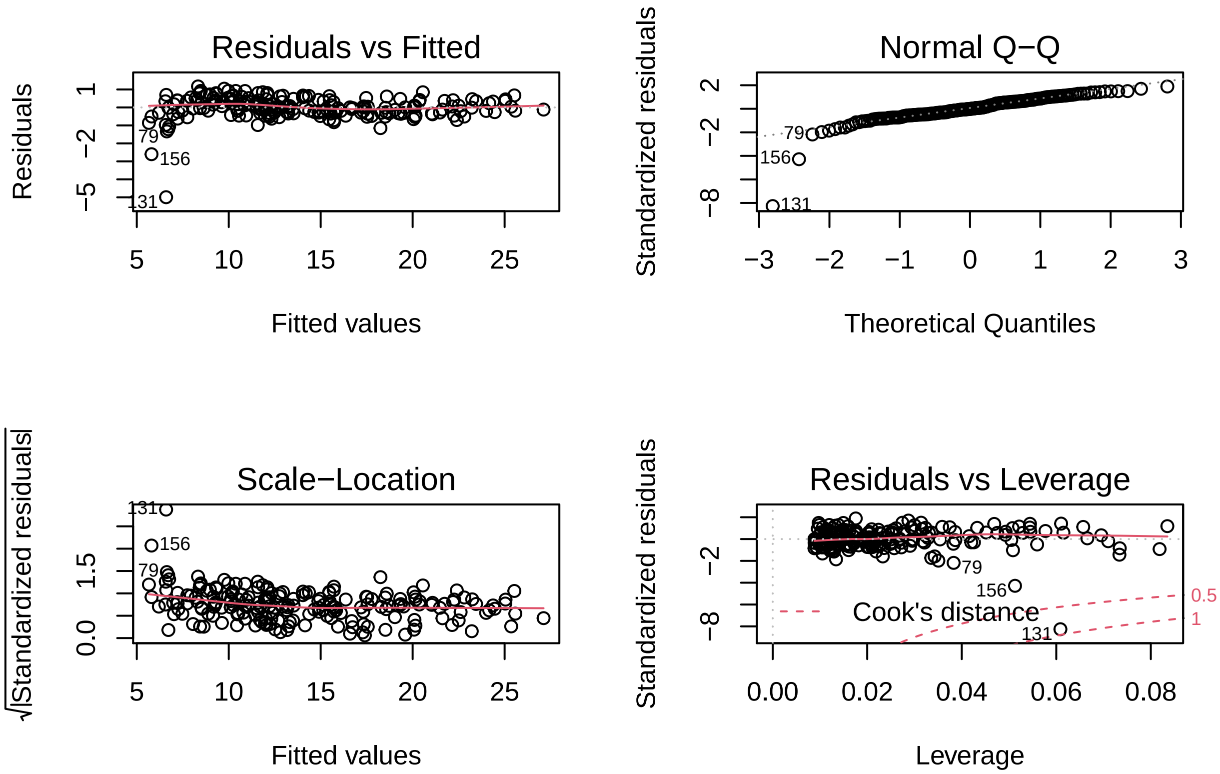
```
m5a = lm(sales ~ TV + I(TV^2) + I(TV^3), data = d)
TV_grid = seq(from = min(d$TV), to = max(d$TV), by = 1)
```

```
m5 = lm(sales ~ TV * radio + I(TV^2), data = d)
summary(m5)
```

```
##
## Call:
## lm(formula = sales ~ TV * radio + I(TV^2), data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9949 -0.2969 -0.0066  0.3798  1.1686
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.137e+00  1.927e-01  26.663  < 2e-16 ***
## TV           5.092e-02  2.232e-03  22.810  < 2e-16 ***
## radio        3.516e-02  5.901e-03   5.959 1.17e-08 ***
## I(TV^2)     -1.097e-04  6.893e-06 -15.920  < 2e-16 ***
## TV:radio     1.077e-03  3.466e-05  31.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6238 on 195 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.9857
## F-statistic:  3432 on 4 and 195 DF,  p-value: < 2.2e-16
```

Comprobamos los residuos:

```r
par(mfrow=c(2,2))
plot(m5)
```

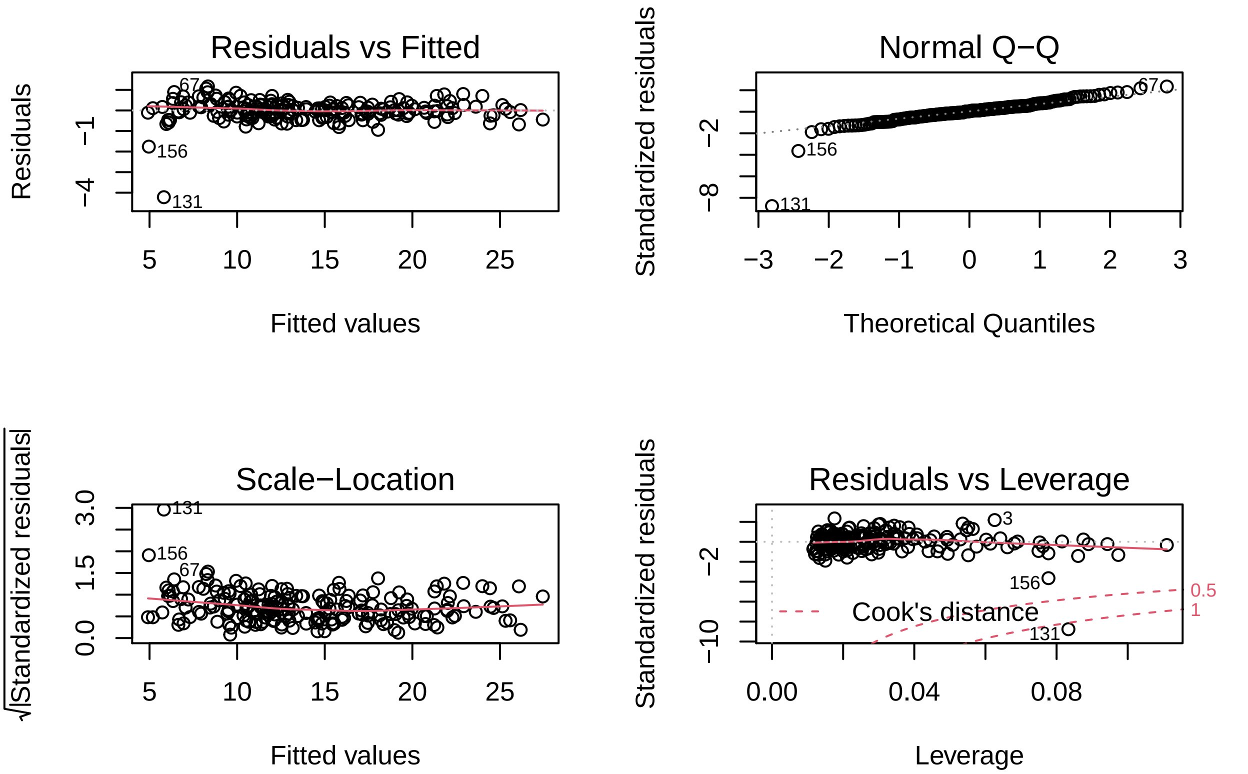


Podemos mejorar un poco mas:

```r
m6 = lm(sales ~ TV * radio + I(TV^2) + I(TV^3), data = d)
summary(m6)
```

```
##
## Call:
## lm(formula = sales ~ TV * radio + I(TV^2) + I(TV^3), data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2184 -0.2106  0.0223  0.2454  1.1677
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.061e+00  1.871e-01  21.709  < 2e-16 ***
## TV           8.998e-02  4.193e-03  21.458  < 2e-16 ***
## radio        4.206e-02  4.801e-03   8.761 9.63e-16 ***
## I(TV^2)     -4.327e-04  3.180e-05 -13.604  < 2e-16 ***
## I(TV^3)      7.278e-07  7.058e-08  10.312  < 2e-16 ***
## TV:radio     1.044e-03  2.811e-05  37.129  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5026 on 194 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9907
## F-statistic:  4250 on 5 and 194 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(m6)
```



Podemos responder a las preguntas:

- ¿Hay relación entre el gasto en publicidad y las ventas?

Podemos utilizar el contraste general de regresión $H_0 : \beta_1 = ... = \beta_5 = 0$, con pvalor $<$ 2.2e-16., luego hay evidencia clara de la relación entre gasto y ventas.

- ¿Es grande esa relación?

Podemos mirar el $R^2 = 0.99$, luego estamos explicando el 99% de la variabilidad de los datos con este modelo.

- ¿Que medios contribuyen a las ventas?

Viendo los contrastes individuales, contribuyen la radio y la TV, su interacción, y terminos polinómicos de la TV.

- ¿Como de grande es el efecto de cada medio?

Es más complicado ver el efecto que en el modelo m3, pero mirando sólo los efectos principales ya que son los más importantes, invertir en TV es más rentable.

- ¿Cual es la precisión de estos valores?

Podemos mirar los *se*:

```r
sqrt(diag(vcov(m6)))
```

```
##   (Intercept)            TV         radio       I(TV^2)       I(TV^3)      TV:radio
## 1.870592e-01 4.193299e-03 4.801434e-03 3.180306e-05 7.057952e-08 2.811053e-05
```

O los intervalos de confianza:

```r
confint(m6)
```

```
##                     2.5 %        97.5 %
## (Intercept)  3.692029e+00  4.429891e+00
## TV           8.171065e-02  9.825127e-02
## radio        3.259386e-02  5.153329e-02
## I(TV^2)     -4.953810e-04 -3.699327e-04
## I(TV^3)      5.886137e-07  8.670171e-07
## TV:radio     9.882675e-04  1.099150e-03
```

- ¿Como de precisas son las predicciones?

Si miramos la predicción del valor medio:

```r
xp = data.frame(TV = 50, radio = 40, newspaper = 60)
exp(predict(m3, newdata = xp, level = 0.95, interval="confidence"))
```

```
##        fit      lwr      upr
## 1 11.11314 10.57021 11.68395
```

```r
predict(m6, newdata = xp, level = 0.95, interval="confidence")
```

```
##       fit      lwr      upr
## 1 11.3393 11.16412 11.51449
```