Práctica 6: Influencia y covariables categóricas

Módulo de Modelos Lineales.
Máster de Bioestadística, Universitat de València.

Miguel A. Martinez-Beneito

Tareas

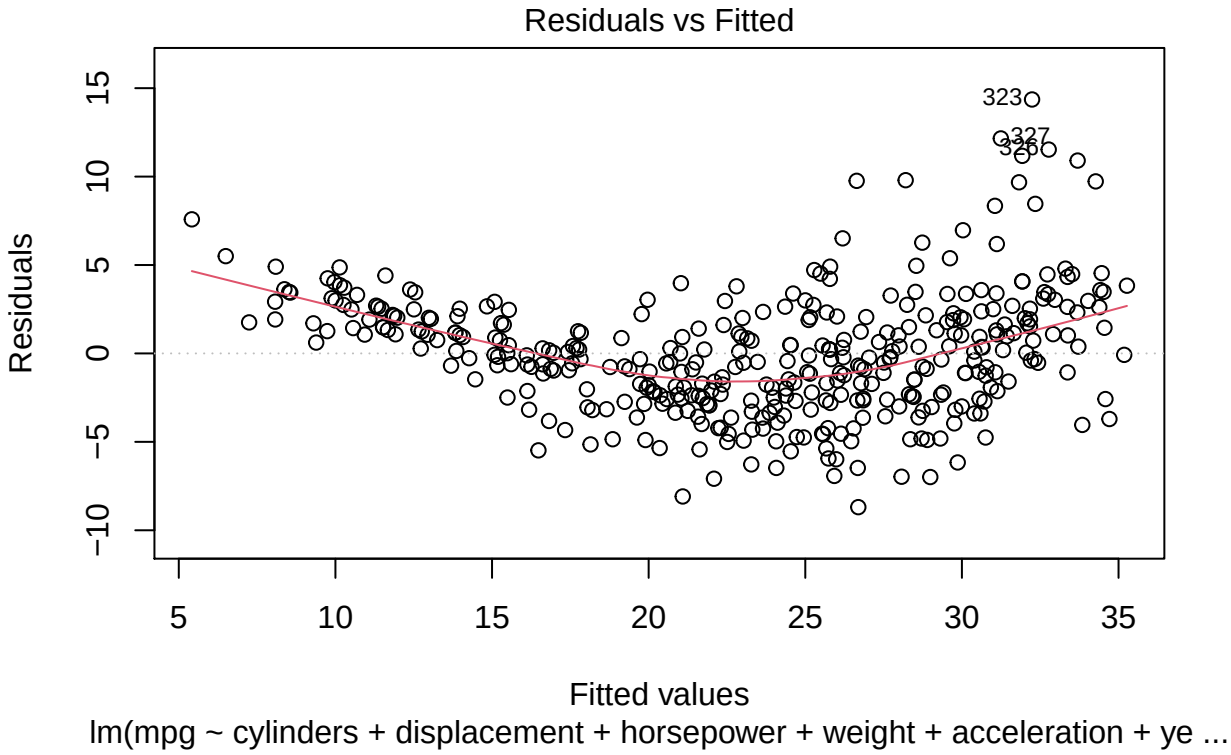
1. Para el modelo que ajustaste en la Tarea 1 de la Práctica 4 (modelo de regresión lineal múltiple para `mpg` en función de todas las variables cuantitativas del banco de datos)

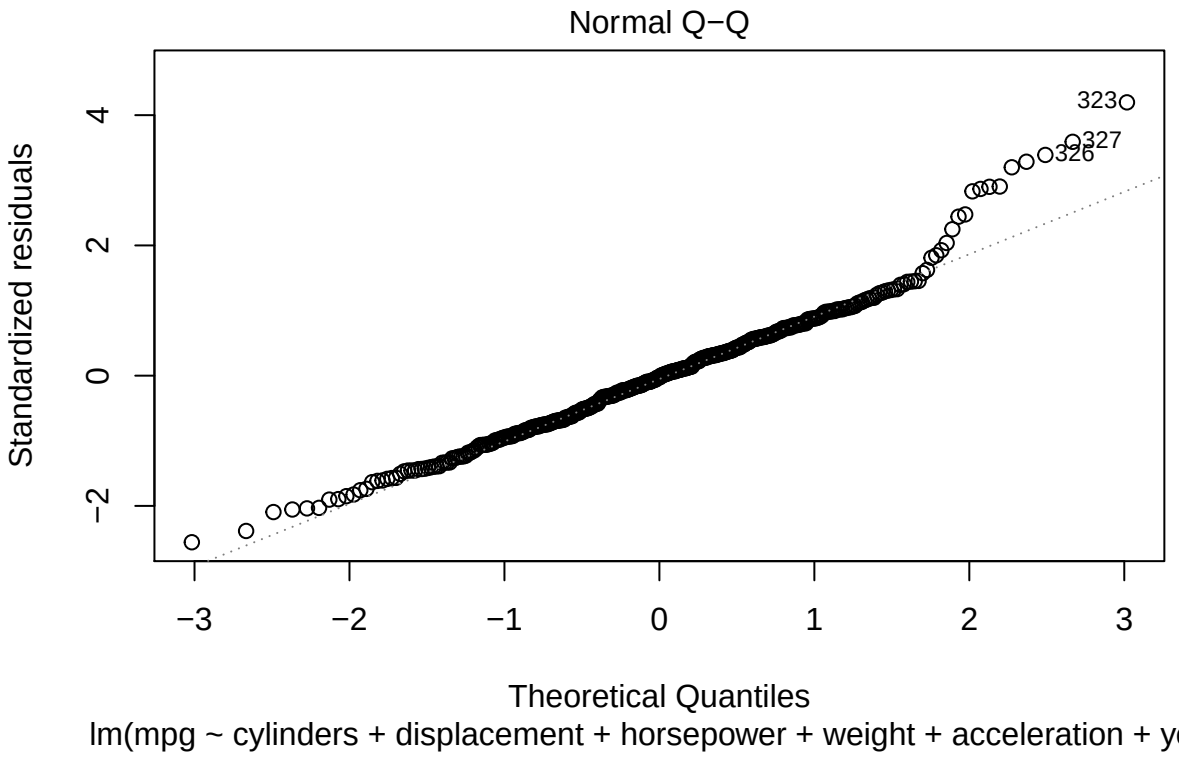
- ¿Observas algún valor particularmente influyente para el modelo?

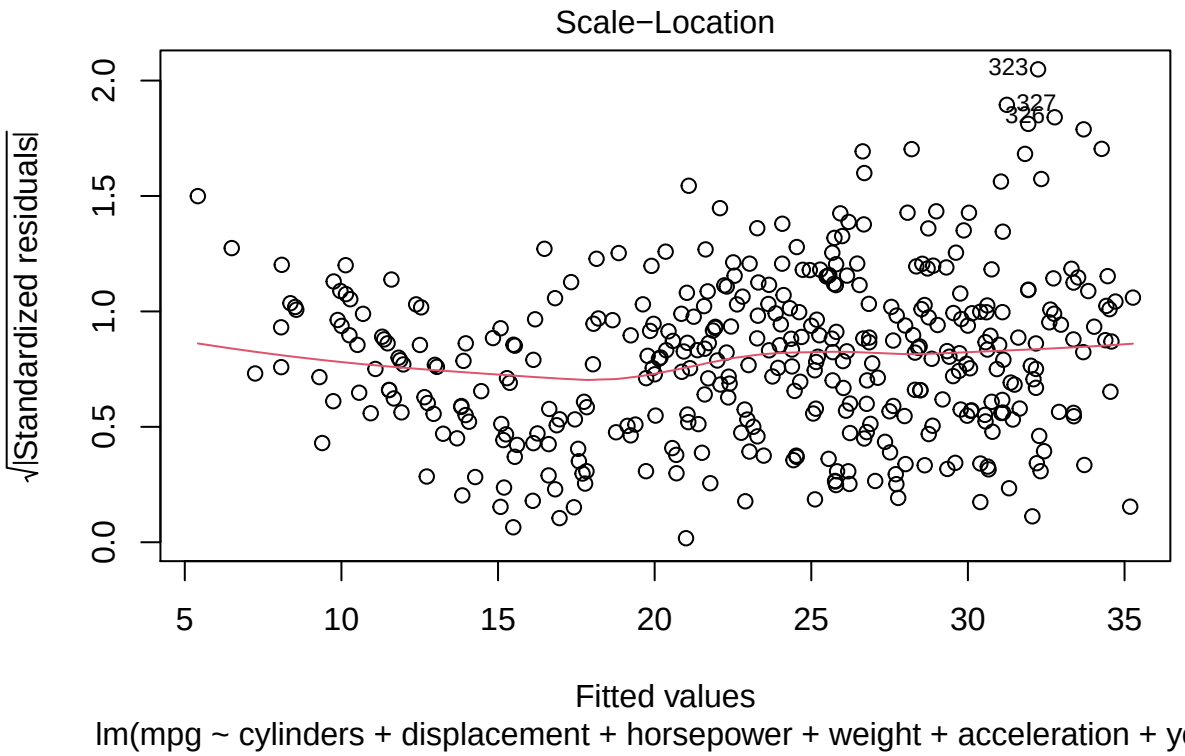
```
data(Auto, package = "ISLR")

modelo <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year,
    data = Auto)
summary(modelo)
```

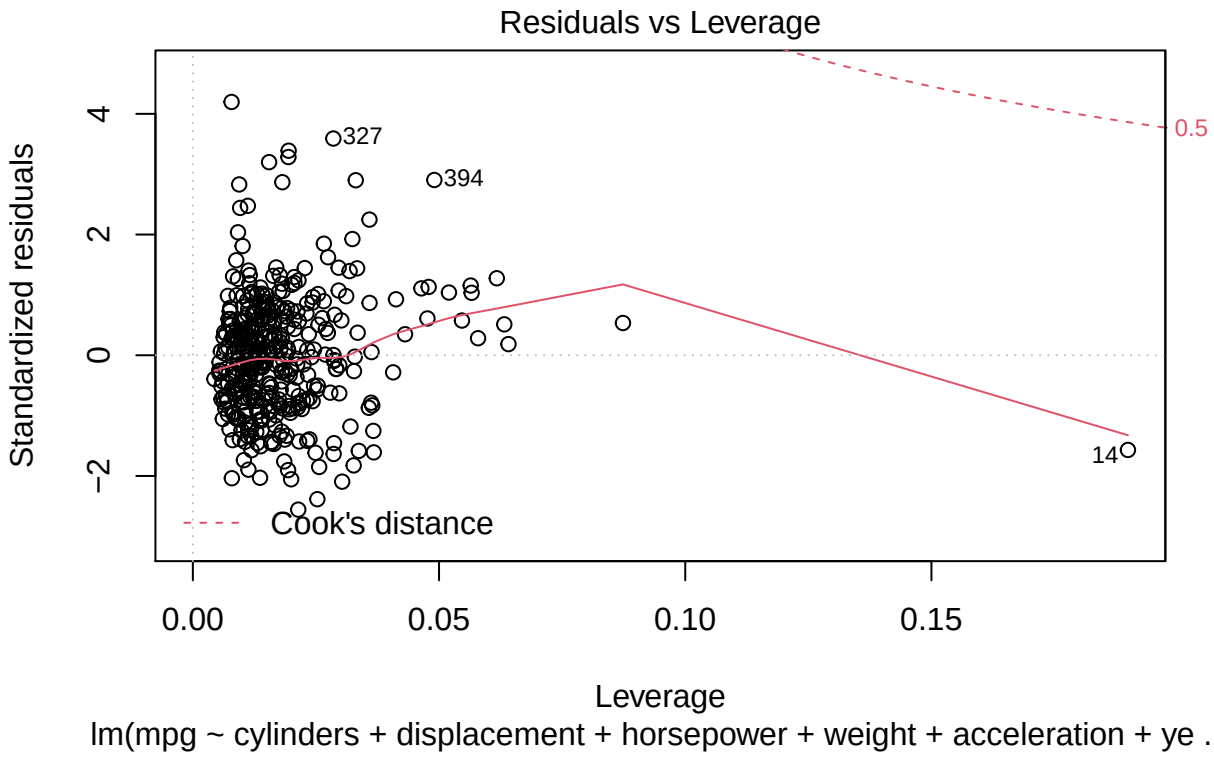
```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6927 -2.3864 -0.0801  2.0291 14.3607
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.454e+01  4.764e+00  -3.051  0.00244 **
## cylinders    -3.299e-01  3.321e-01  -0.993  0.32122
## displacement  7.678e-03  7.358e-03   1.044  0.29733
## horsepower   -3.914e-04  1.384e-02  -0.028  0.97745
## weight       -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
## acceleration  8.527e-02  1.020e-01   0.836  0.40383
## year          7.534e-01  5.262e-02  14.318  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.435 on 385 degrees of freedom
## Multiple R-squared:  0.8093, Adjusted R-squared:  0.8063
## F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

```
plot(modelo)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...
323
327
326
15
10
5
0
−5
−10
5
10
15
20
25
30
35


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...
323
327
326
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
323
327
326
5
10
15
20
25
30
35
Fitted values
lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...




```
# No parece haber ningún individuo con distancia de Cook particularmente destacada por
# lo que no observamos ningún coche particulamente influyente sobre el modelo que
# hemos ajustado.
```

- ¿Los valores más influyentes que observas tienen un `mpg` inferior o superior al que se esperaría de acuerdo con el modelo?

```
# Los coches más influyentes para el modelo son el 14, 327 y 394. Los dos últimos
# recorren más millas por galón de lo que predice el modelo (residuo positivo), al
# contrario que el número 14.
```

- ¿Observas la existencia de algún vehículo potencialmente influyente en el modelo de regresión ajustado?

```
# El modelo potencialmente más influyente sobre la recta de regresión será el 14
# (leverage alto), aunque no lo es tanto porque su residuo es pequeño (su variable
# respuesta está en consonancia con lo que dice el modelo).
```

- El modelo estudiado tiene un buen número de covariable no significativas ¿Crees que éstas pueden alterar de alguna manera el análisis que acabamos de hacer? ¿De qué forma?

```
# El quitar las variables no significativas del modelo no cambiaría apenas la recta de
# regresión por tanto dichas variables de más no afectan tampoco a los residuos del
# modelo. Sin embargo dichas variables sí que influirán de forma importante sobre los
# leverages de los distintos coches, que al fin y al cabo dependían sólo de las
# covariables, por tanto habría sido aconsejable quitar primero las covariables no
# significativas del modelo antes de analizar la influencia de sus observaciones.
```

2. Para el banco de datos `Auto`, considera el modelo sin interacción que estimaras más adecuado de los que ajustaras en la práctica 4.

- Incluye ahora en el modelo de regresión `origin` como variable categórica.

```
# Modelo original
modelo2 <- lm(mpg ~ weight + year, data = Auto)
summary(modelo2)
```

```
##
## Call:
## lm(formula = mpg ~ weight + year, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8505 -2.3014 -0.1167  2.0367 14.3555
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.435e+01  4.007e+00  -3.581 0.000386 ***
## weight      -6.632e-03  2.146e-04 -30.911  < 2e-16 ***
## year         7.573e-01  4.947e-02  15.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.427 on 389 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.8072
## F-statistic: 819.5 on 2 and 389 DF,  p-value: < 2.2e-16
```

```
# Pongo etiquetas a las categorías de origin
Auto$origin <- factor(Auto$origin, labels = c("EEUU", "Europa", "Japón"))
# Modelo con origin
modelo3 <- lm(mpg ~ weight + year + origin, data = Auto)
summary(modelo3)
```

```
##
## Call:
## lm(formula = mpg ~ weight + year + origin, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6025 -2.1132 -0.0206  1.7617 13.5261
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.831e+01  4.017e+00  -4.557 6.96e-06 ***
## weight       -5.887e-03  2.599e-04 -22.647  < 2e-16 ***
## year          7.698e-01  4.867e-02  15.818  < 2e-16 ***
## originEuropa  1.976e+00  5.180e-01   3.815 0.000158 ***
## originJapón   2.215e+00  5.188e-01   4.268 2.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.337 on 387 degrees of freedom
## Multiple R-squared:  0.819,  Adjusted R-squared:  0.8172
## F-statistic: 437.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
# El origen del coche parece tener un efecto significativo, ya que 2 de sus efectos
# son significativos. Los coches japones son más eficientes y los que menos los
```

```
# norteamericanos.
```

- ¿Consideras que hay diferencias significativas en cuanto a `mpg` entre los coches europeos y japoneses?

```
# Pongo como referencia los coches europeos para poder comparar los japones frente a
# éstos
Auto$origin2 <- relevel(Auto$origin, ref = "Europa")
modelo4 <- lm(mpg ~ weight + year + origin2, data = Auto)
summary(modelo4)
```

```
##
## Call:
## lm(formula = mpg ~ weight + year + origin2, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6025 -2.1132 -0.0206  1.7617 13.5261
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.633e+01  3.927e+00  -4.158 3.95e-05 ***
## weight       -5.887e-03  2.599e-04 -22.647  < 2e-16 ***
## year          7.698e-01  4.867e-02  15.818  < 2e-16 ***
## origin2EEUU  -1.976e+00  5.180e-01  -3.815 0.000158 ***
## origin2Japón  2.382e-01  5.591e-01   0.426 0.670275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.337 on 387 degrees of freedom
## Multiple R-squared:  0.819,  Adjusted R-squared:  0.8172
## F-statistic: 437.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
# No encontramos diferencias significativas entre los coches europeos y japoneses
```

- Valora la presencia de interacción entre el origen de los vehículos y las variables que estuvieran ya anteriormente en el modelo. Interpreta los resultados obtenidos: ¿Qué región ha tenido una mejor evolución temporal en cuanto al consumo de sus coches? ¿En que región el peso de los coches tiene un efecto más importante en el consumo?

```
modelo5 <- lm(mpg ~ (weight + year) * origin2, data = Auto)
summary(modelo5)
```

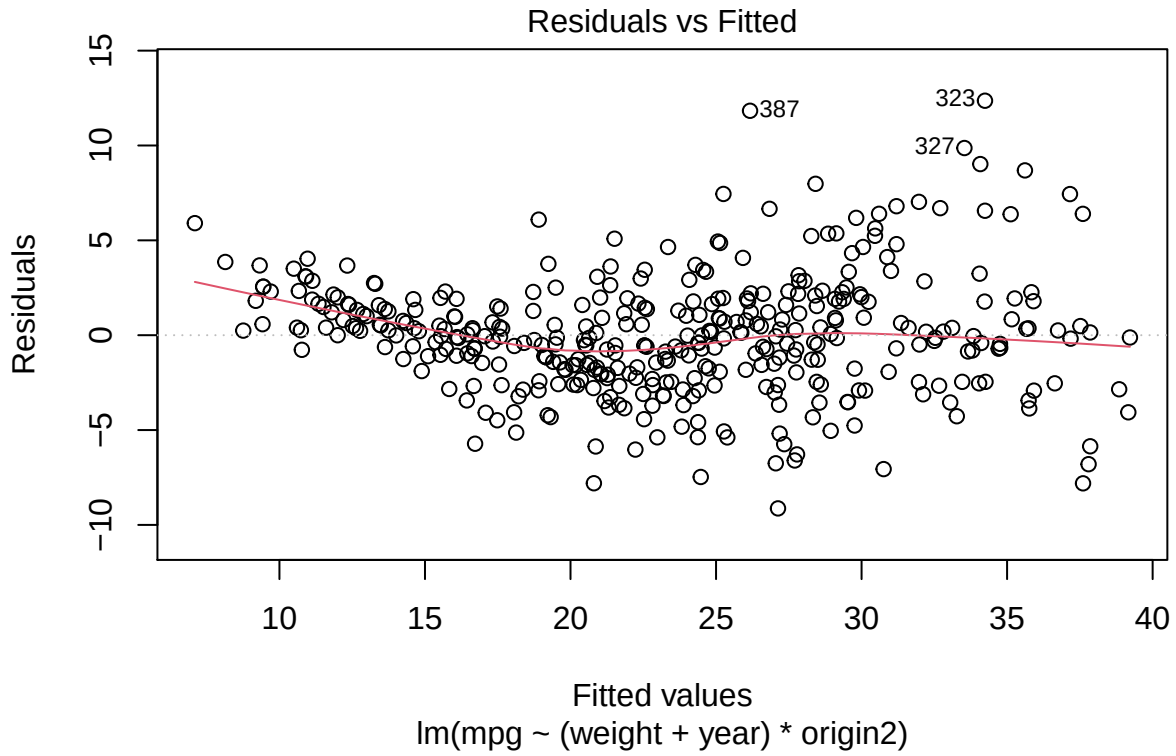
```
##
## Call:
## lm(formula = mpg ~ (weight + year) * origin2, data = Auto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.129 -1.899 -0.049  1.764 12.360
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.163e+01  8.494e+00  -4.901 1.41e-06 ***
## weight              -8.316e-03  7.922e-04 -10.498  < 2e-16 ***
## year                 1.182e+00  1.138e-01  10.387  < 2e-16 ***
## origin2EEUU          3.219e+01  9.855e+00   3.267  0.00119 **
```

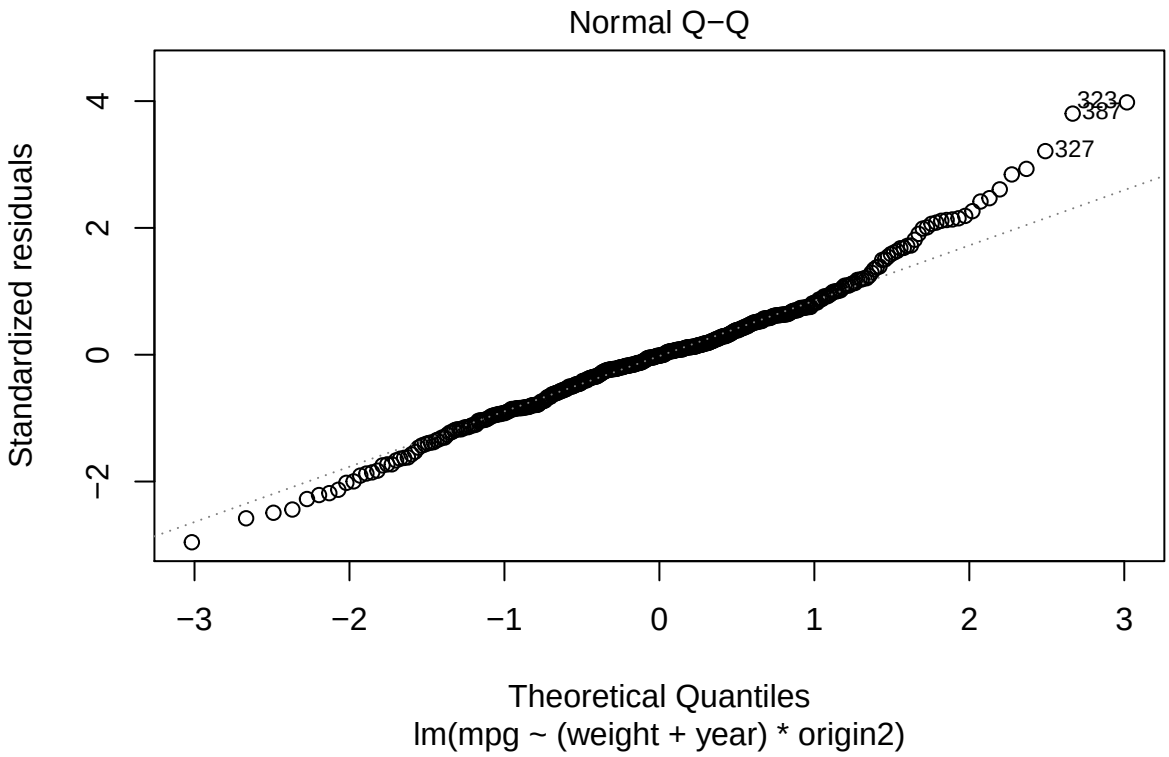
```
## origin2Japón          2.010e+01  1.156e+01   1.739  0.08280 .
## weight:origin2EEUU    2.663e-03  8.390e-04   3.174  0.00162 **
## weight:origin2Japón -2.916e-03  1.363e-03  -2.139  0.03310 *
## year:origin2EEUU     -5.402e-01  1.287e-01  -4.197 3.36e-05 ***
## year:origin2Japón    -1.889e-01  1.498e-01  -1.261  0.20810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.138 on 383 degrees of freedom
## Multiple R-squared:  0.8417, Adjusted R-squared:  0.8384
## F-statistic: 254.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

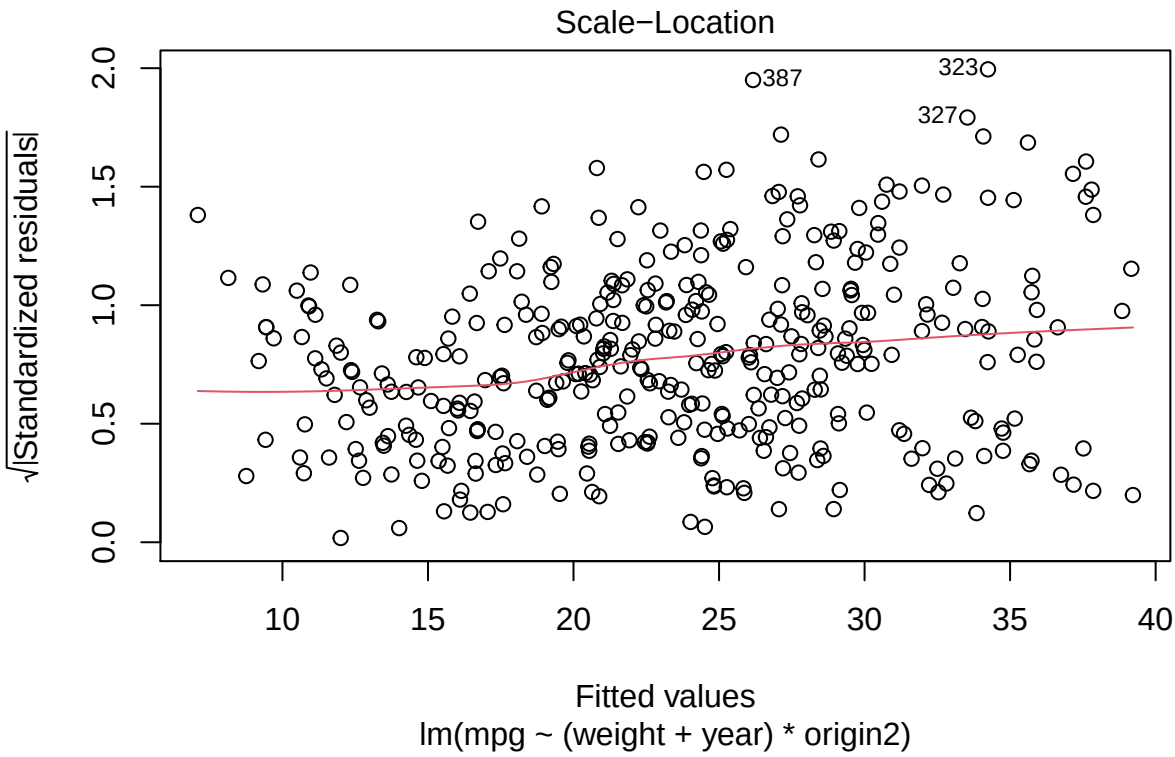
```
# Parece observarse interacción entre el origen del coche y el peso y año de
# fabricación.  Los coches europeos han tenido una mejor evolución temporal en cuanto
# al consumo, la pendiente de la recta de regresión que explica las millas por galón
# frente al año de construcción es más positiva en los coches europeos, aunque las
# diferencias con los coches japoneses no es significativa.  Los coches americanos son
# aquellos donde el peso tiene un efecto más importante sobre el consumo ya que los
# coches que pertenecen a dicha región presentan una recta más inclinada en función
# del peso.
```

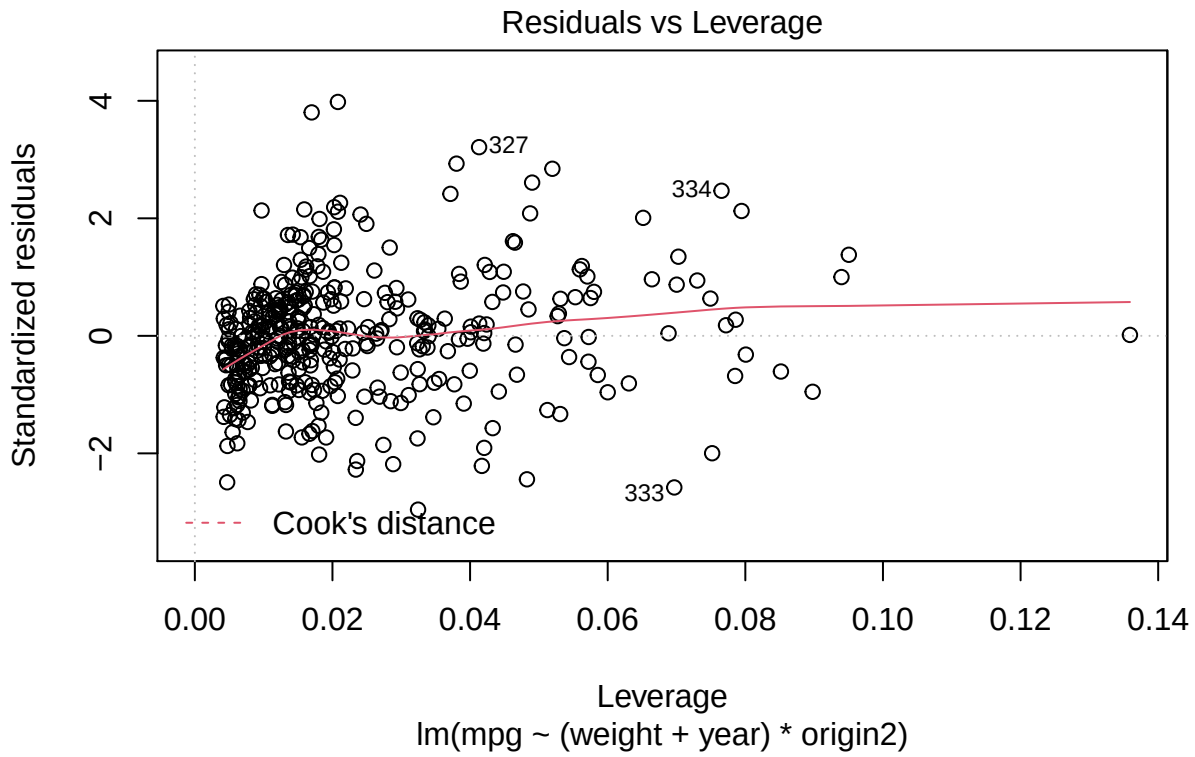
- Sobre este último modelo que has ajustado ¿Qué vehículos consideras que han podido tener más influencia sobre el ajuste de la recta de regresión?

```
plot(modelo5)
```




Normal Q−Q
Standardized residuals
4
2
0
−2
323
387
327
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ (weight + year) * origin2)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
387
323
327
10
15
20
25
30
35
40
Fitted values
lm(mpg ~ (weight + year) * origin2)




```
# Los vehículos 327, 333 y 334. El vehículo 14 ya no tiene un leverage alto.
```

- Por último, sobre el último modelo que has ajustado, incluye como variable explicativa el número de cilindros de cada coche junto a su interacción con la región de procedencia del vehículo ¿Consideras que la interacción con la región de fabricación tienen un efecto significativo en cuanto a la explicación de `mpg`?

```
# Modelo con número de cilindros e interacción
modelo6 <- lm(mpg ~ (weight + year + cylinders) * origin2, data = Auto)
summary(modelo6)
```

```
##
## Call:
## lm(formula = mpg ~ (weight + year + cylinders) * origin2, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8331 -1.8517 -0.0766  1.6553 12.3975
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -40.194801   8.413706  -4.777 2.54e-06 ***
## weight                 -0.007453   0.000939  -7.937 2.35e-14 ***
## year                    1.220027   0.114444  10.661  < 2e-16 ***
## cylinders              -1.536000   0.929057  -1.653 0.099098 .
## origin2EEUU            34.282562   9.827639   3.488 0.000543 ***
## origin2Japón           16.909053  11.455614   1.476 0.140759
```

```
## weight:origin2EEUU       0.003076   0.001077   2.856 0.004526 **
## weight:origin2Japón     -0.004978   0.001565  -3.180 0.001592 **
## year:origin2EEUU        -0.620206   0.129705  -4.782 2.49e-06 ***
## year:origin2Japón       -0.240467   0.149536  -1.608 0.108648
## cylinders:origin2EEUU    0.798003   0.965075   0.827 0.408823
## cylinders:origin2Japón   2.874421   1.152287   2.495 0.013036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.092 on 380 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8431
## F-statistic:   192 on 11 and 380 DF,  p-value: < 2.2e-16
```

```
# Esta salida no nos informa sobre la significatividad de la interacción
# cilindros:origen de manera global

# Modelo con número de cilindros y sin interacción
modelo7 <- lm(mpg ~ (weight + year) * origin2 + cylinders, data = Auto)
anova(modelo7, modelo6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ (weight + year) * origin2 + cylinders
## Model 2: mpg ~ (weight + year + cylinders) * origin2
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    382 3721.2
## 2    380 3632.2  2    89.067 4.6591 0.01002 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# El efecto de la interacción sí es significativo
```