

Supuestos de Linealidad

bas, magí

2023-05-11

Regresión Lineal: Requisitos

Todos los derechos reservados a Data Política, en su vídeo encontrareis el doc 'trabajadores.sav':

Font: Data Política (<https://www.youtube.com/watch?v=FGpfhKwsluE&list=WL&index=1>)
(<https://www.youtube.com/watch?v=FGpfhKwsluE&list=WL&index=1>))

Este es un documento genial para repasar los supuestos de linealidad cuando estamos trabajando con nuestro modelo.

```
# Construimos el modelo
formula= data$salario_actual~data$salario_inicial + data$experiencia + data$antiguedad
modelo=lm(formula)
summary(modelo)
```

```
##
## Call:
## lm(formula = formula)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -32736  -3965  -1214    2458   46474
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.027e+04  2.960e+03  -3.469 0.000571 ***
## data$salario_inicial  1.927e+00  4.437e-02  43.435 < 2e-16 ***
## data$experiencia    -2.251e+01  3.339e+00  -6.742 4.59e-11 ***
## data$antiguedad      1.732e+02  3.468e+01   4.995 8.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7586 on 470 degrees of freedom
## Multiple R-squared:  0.8039, Adjusted R-squared:  0.8026
## F-statistic: 642.2 on 3 and 470 DF, p-value: < 2.2e-16
```

Preguntas para poder validar un modelo:

- $p\text{-valor} < 0.05 \Rightarrow$ Modelo válido ●
- R^2 Ajustado \Rightarrow Explica un 80% de la variabilidad de y
- Cada una de las variables independientes aporta al model

Supuestos

Para verificar un modelo lineal necesitamos verificar que se cumplan los siguientes supuestos:

- Linealidad
- Normalidad de residuos
- Homocedasticidad (Homogeneidad de variancias)
- Ausencia de multicolinealidad
- Ausencia de valores influyentes

Linealidad

La linealidad se produce cuando existe relación lineal entre las variables independientes y la variable dependiente.

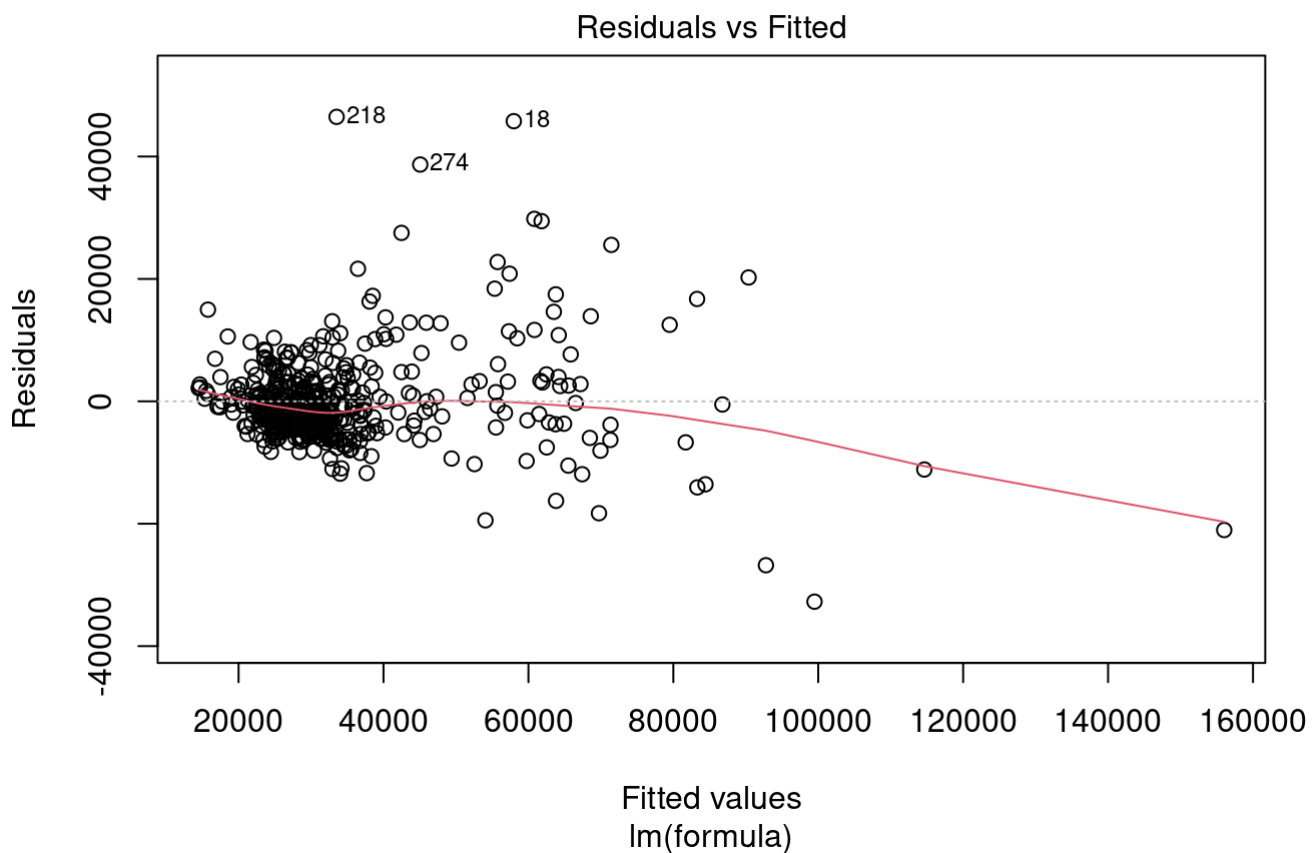
¿Porque es un problema cuando no se cumple?

Porque se pueden considerar variables que no aporten al modelo. También porque pueden existir otras relaciones no lineales que no son vistas previamente.

¿Como detectarlo?

Con verificación gráfica: Diagrama de dispersión entre los valores predichos y los residuos (la línea de tendencia debe ser horizontal) O pruebas de linealidad para cada una de las variables independientes.

```
plot(modelo, 1)
```




Vemos un gráfico de los valores predichos y cada uno de los residuos, la línea roja se aleja de la línea puntual. Solo viendo el gráfico podríamos decir que hay un problema en el supuesto de linealidad, tendríamos que verificar linealidad para cada una de las variables independientes con la dependiente.

```
cor.test(data$salario_actual, data$salario_inicial)
```

```
##
## Pearson's product-moment correlation
##
## data: data$salario_actual and data$salario_inicial
## t = 40.276, df = 472, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8580696 0.8989267
## sample estimates:
##          cor
## 0.8801175
```

Test de Pearson

[(p-value < 0.05) = (H0: No existe correlacion lineal; H1: Existe correlación lineal)]

Relación lineal 


```
cor.test(data$salario_actual, data$experiencia)
```

```
##
## Pearson's product-moment correlation
##
## data: data$salario_actual and data$experiencia
## t = -2.1277, df = 472, p-value = 0.03388
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.185900660 -0.007466824
## sample estimates:
##          cor
## -0.09746693
```



```
cor.test(data$salario_actual, data$antiguedad)
```

```
##
## Pearson's product-moment correlation
##
## data: data$salario_actual and data$antiguedad
## t = 1.8334, df = 472, p-value = 0.06737
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.006018969 0.172848789
## sample estimates:
##          cor
## 0.08409227
```

 No cumple linealidad con la variable independiente, podriamos excluirla o analizar porque no es lineal.

Normalidad de residuos

La normalidad de residuos existe cuando los residuos (no tipificados) del modelo no siguen una distribución normal.

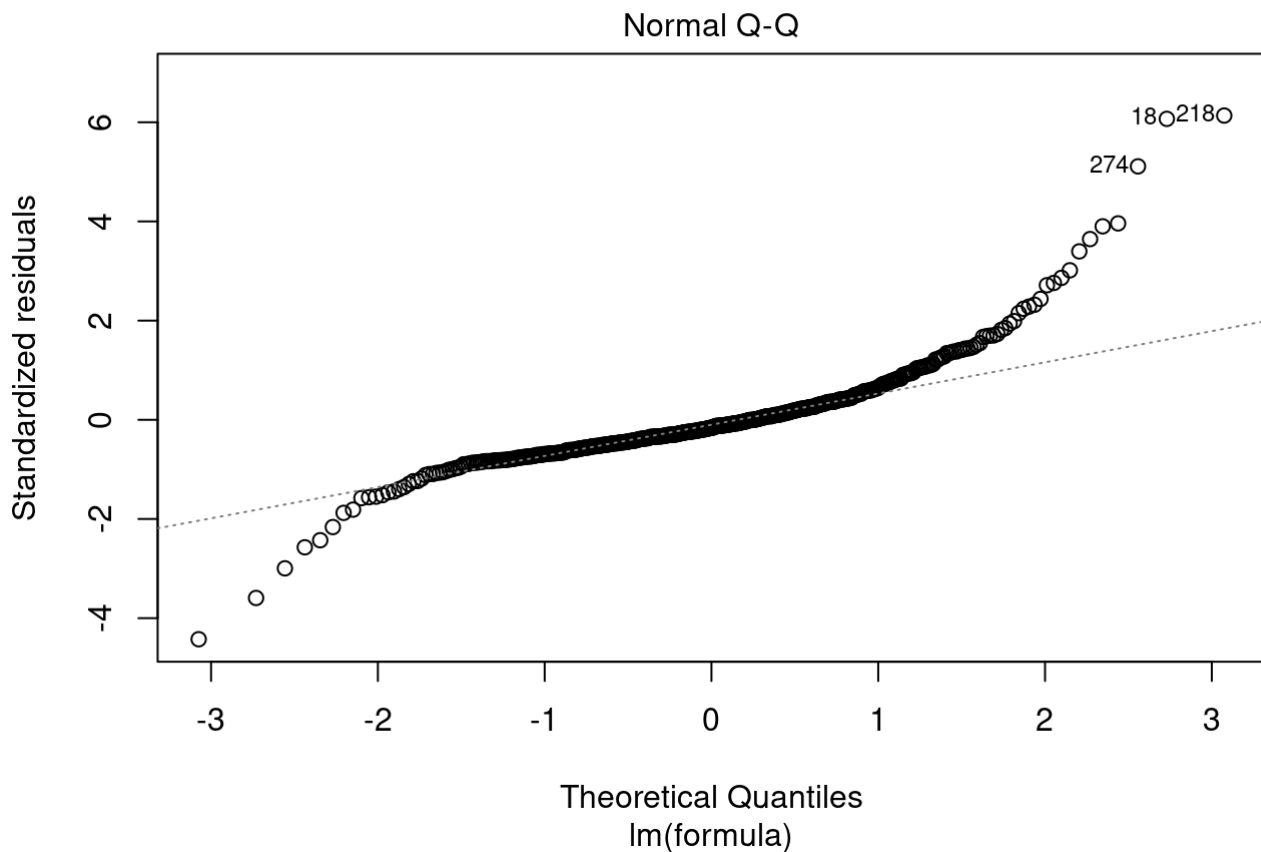
¿Porque es un problema cuando no se cumple?

Porque no se podrían aplicar pruebas de validación global del modelo (ANOVA). Estas tienen como principal requerimiento que exista normalidad.

¿Como detectarlo?

Con un qqplot o realizar una prueba de normalidad (Shapiro Test)

```
# Gráfico
plot(modelo, 2)
```



En un supuesto de linealidad esperamos ver como todos los residuos siguen la linea daigonal. En este casa, y aunque la mayoría de puntos lo cumplen tenemos una dispersión hacia el final e inicio del gráfico, para una cuantificación del supuesto de linealidad ejecutamos el test.

```
# Test de normalidad
shapiro.test(modelo$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.84851, p-value < 2.2e-16
```

Shapiro Test*[H0: Normal; H1: No normal]*

p-value < 0.05 => ● H1

Homocedasticidad de residuos

También llamada homogeneidad de varianzas, cuando la varianza de los residuos es constante. Si no lo es, diremos entonces que el modelo es heterocedástico.

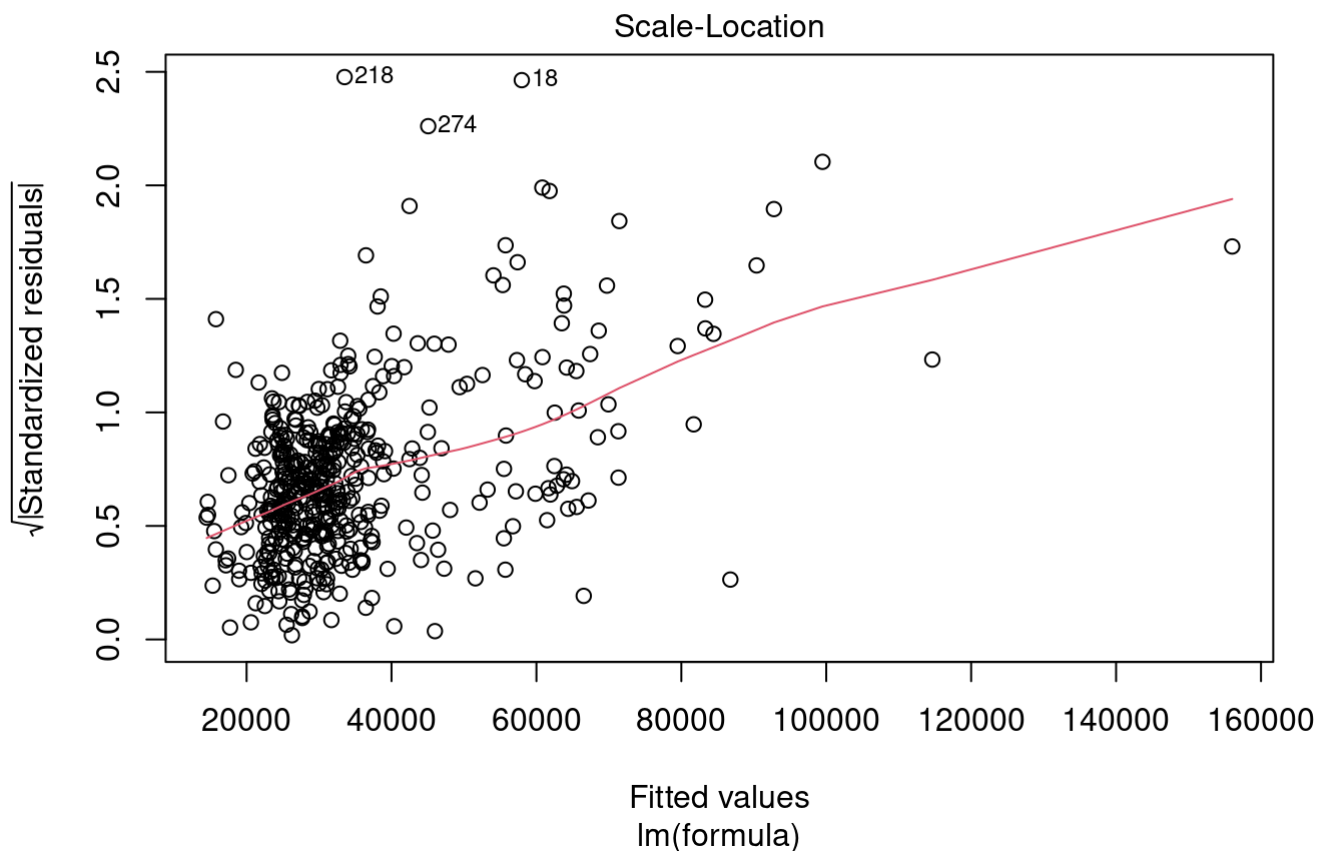
¿Porque es un problema cuando no se cumple?

Si existe un patrón puede ser que el modelo no funcione bien. Se asume que el error del modelo de regresión no afecta a la varianza o dispersión de la estimación.

¿Como detectarlo?

- Grafico de dispersión teniendo los residuos estandarizados en el eje Y, y los valores pronosticados en el eje X. (la línea roja debe tender a ser horizontal)
- Test Breusch Pagan [H0: Existe homocedasticidad]

```
# Grafico
plot(modelo, 3)
```



Se puede ver claramente como no estamos frente a un modelo homocedástico, ya que ni los puntos están uniformemente distribuidos ni la línea roja es horizontal.

Para corroborar la no homocedasticidad:

```
# Test BP
if(!require('car')){install.packages("car")}
```

```
## Loading required package: car
```


```
## Loading required package: carData
```

```
library(car)
ncvTest(modelo)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 315.8062, Df = 1, p = < 2.22e-16
```

Non-constant Variance Score Test

[H0: Homocedasticidad; H1: No homocedasticidad]

p-valor < 0.05 => H1 

Ausencia de multicolinealidad

Se produce cuando existe una fuerte o total correlación entre las variables independientes (x).

¿Por que es un problema cuando se presenta?

Cuando la colinealidad es alta produce coeficientes muy inestables en la ecuación. En otras palabras, los efectos atribuidos a las variables independientes pueden ser engañosos.

¿Como detectarlo?

Analizando el estadístico VIF(factor de inflación de varianza). Cuando VIF >5, hay problemas de multicolinealidad

```
# Test VIF
if(!require('DescTools')){install.packages("DescTools")}
```

```
## Loading required package: DescTools
```

```
##
## Attaching package: 'DescTools'
```

```
## The following object is masked from 'package:car':
##
##      Recode
```

```
library(DescTools)
VIF(modelo)
```

```
## data$salario_inicial    data$experiencia    data$antiguedad
##                1.002439                1.002056                1.000405
```

Todos los VIF son menores a 5, es decir, no tenemos un escenario de multicolinealidad. ●

Identificación de valores influyentes

Una observación influyente se define como una observación que se diferencia marcadamente del conjunto de datos y tiene una gran influencia en el resultado del modelo, es decir, que no solo son outliers.

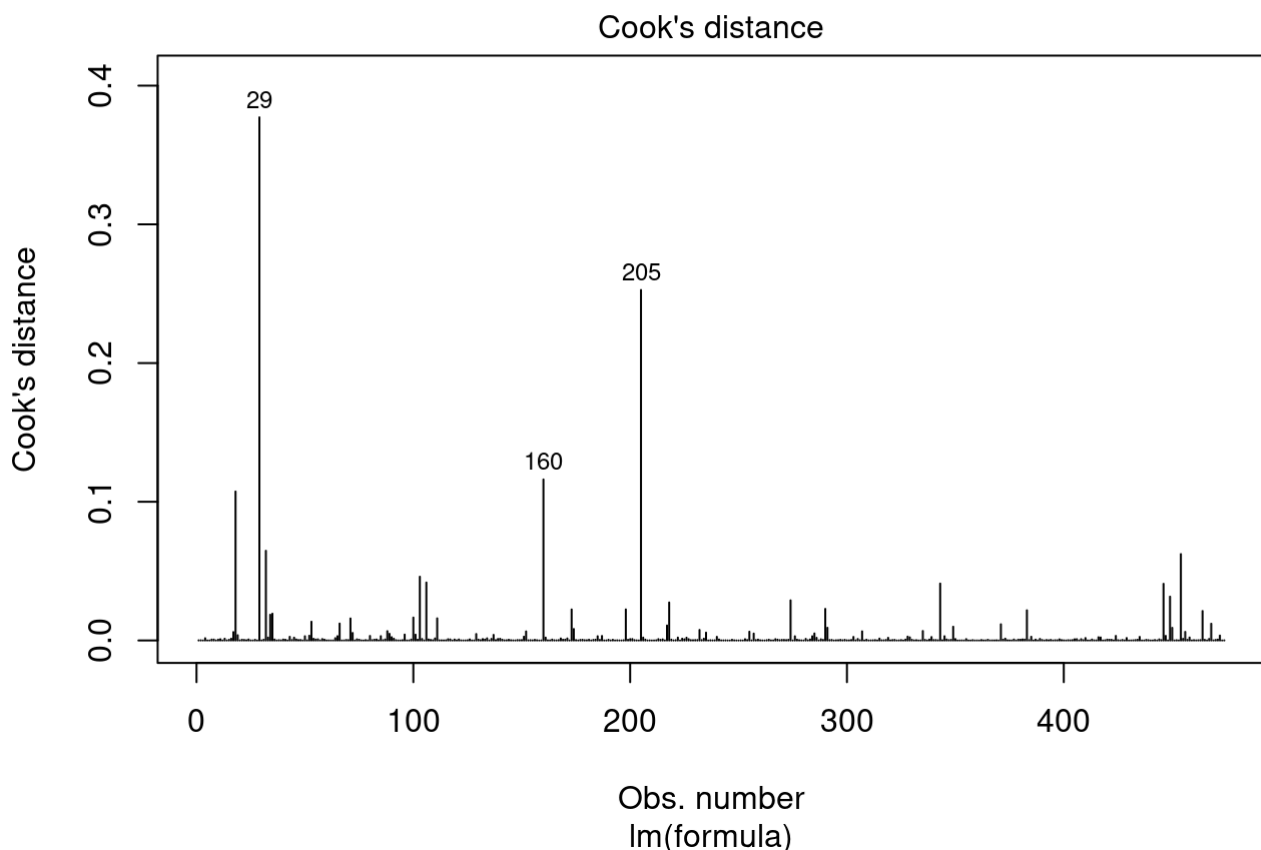
¿Por que son un problema?

Porque afectan los coeficientes de la ecuación y generan errores de predicción

¿Como detectarlos?

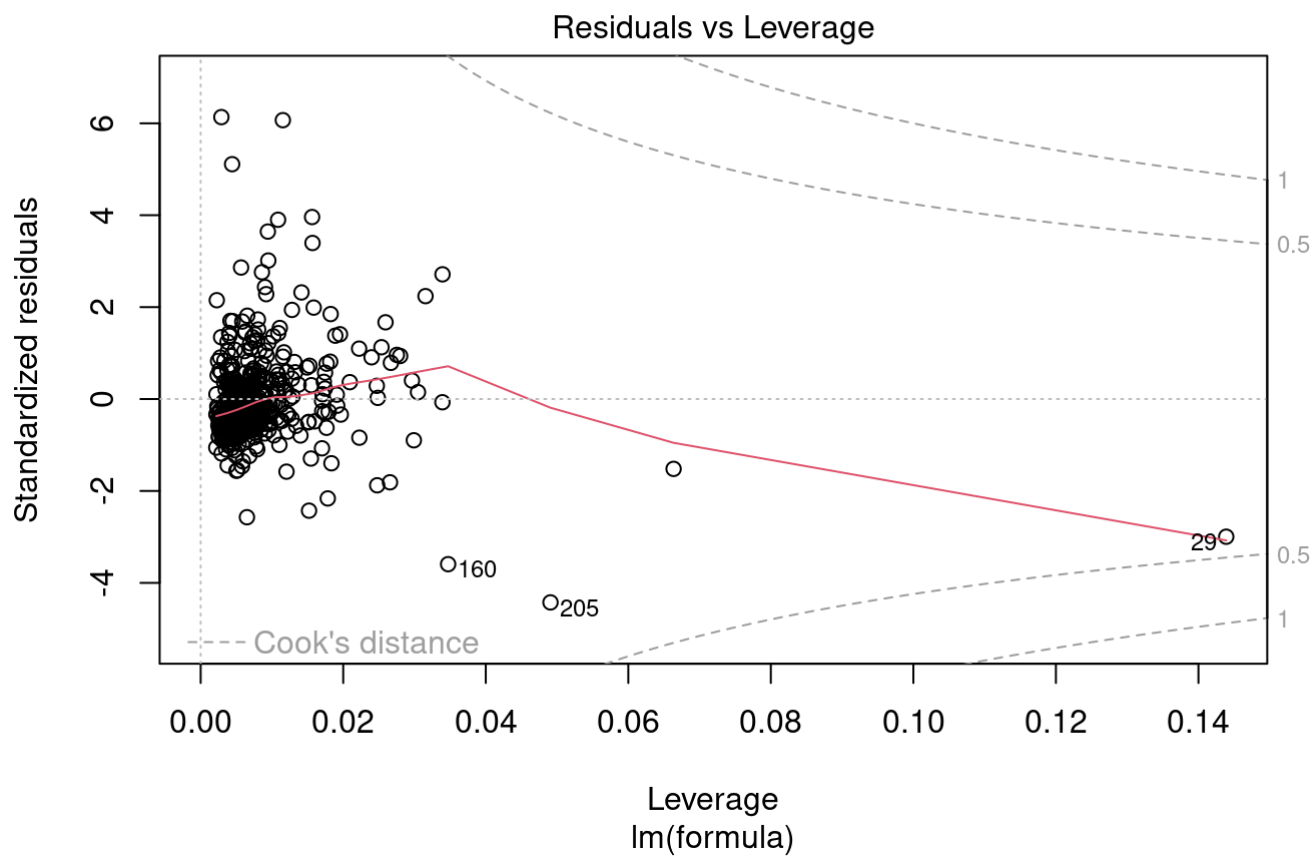
Se utilizan medidas de influencia, entre las que resalta la distancia de Cook. LA distancia de Cook indica que un caso es un valor influyente cuando $DCook \geq 1$

```
# Grafico
plot(modelo, 4)
```



Es importante ver la escala del gráfico, ya que, en este caso, ninguno pasa del valor 1.

```
plot(modelo, 5)
```



No hay ninguna observación que pase de las regiones del gráfico.

```
# Distancia de cook
data$cook=cooks.distance(modelo)
which(data$cook > 1)
```

```
## integer(0)
```