Regresión Múltiple

Jorge López Díaz

2025-11-18

Descripción y objetivos

Se dispone de información (Belsley, Kuh y Welsch,1980) sobre promedios de 1960-1970 (para eliminar el ciclo económico u otras fluctuaciones a corto plazo) de 50 países sobre las siguientes variables:
- dpi es la renta disponible per cápita en dólares;
- ddpi es la tasa de variación porcentual de la renta disponible per cápita;
- sr es el ahorro personal agregado dividido por la renta disponible.
- pop15 es el porcentaje de población menor de 15 años.
- pop75 es el porcentaje de población mayor de 75 años. La información está disponible en el conjunto de datos savings del paquete faraway. Se solicita utilizar la información anterior para construir un modelo de regresión lineal que permita predecir la capacidad de ahorra personal de la población (variable respuesta sr) basandose en un análisis estadístico exhaustivo.

FASE I

```
if (!require(faraway )) install.packages("faraway")
if (!require(ggplot2 )) install.packages("ggplot2")
if (!require(car )) install.packages("car")
if (!require(corrplot )) install.packages("corrplot")
if (!require(MASS )) install.packages("MASS")
```

Importar y preparar las variables. Se carga la librería que almacena los datos.

```
library(faraway)
```

Carga de datos y preparación de variables.

```
sv <- data.frame(savings[,1:3], dpi=(savings[,4])/1000, ddpi=savings[,5])
nrow(sv)
```

```
## [1] 50
```

```
names <- rownames(sv)
```

Descripción de variables.

```
summary(sv)
```

```
##        sr             pop15           pop75            dpi
##  Min.   : 0.600   Min.   :21.44   Min.   :0.560   Min.   :0.08894
##  1st Qu.: 6.970   1st Qu.:26.21   1st Qu.:1.125   1st Qu.:0.28821
##  Median :10.510   Median :32.58   Median :2.175   Median :0.69566
##  Mean   : 9.671   Mean   :35.09   Mean   :2.293   Mean   :1.10676
##  3rd Qu.:12.617   3rd Qu.:44.06   3rd Qu.:3.325   3rd Qu.:1.79562
##  Max.   :21.100   Max.   :47.64   Max.   :4.700   Max.   :4.00189
##       ddpi
##  Min.   : 0.220
##  1st Qu.: 2.002
##  Median : 3.000
##  Mean   : 3.758
##  3rd Qu.: 4.478
##  Max.   :16.710
```

Los datos de ingresos (dpi) y su tasa de crecimiento (ddpi) presentan una marcada asimetría positiva, mientras que las otras variables muestran distribuciones más centrales.

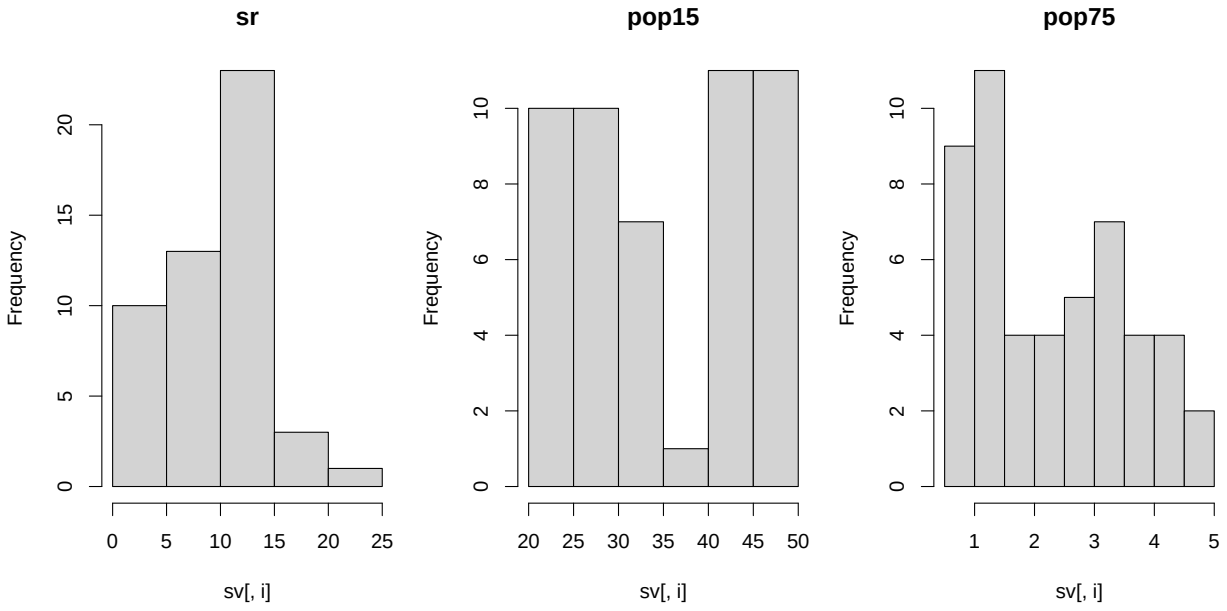
```
View(sv)
```

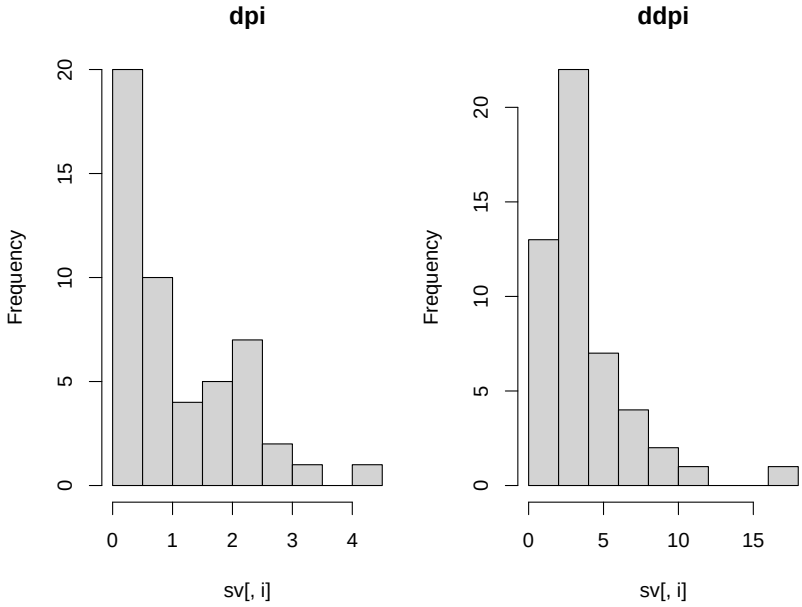
No existen valores ausentes.

FASE II

Análisis descriptivo de las variables, estudio de valores ausentes o atípicos y estudio de la correlación. *Histogramas*

```
par(mfrow = c(2,3),
    mar = c(4,4,3,1),
    cex = 1.2)
for(i in 1:5) hist(sv[,i],main=names(sv)[i])
```





sr: Distribución con sesgo a la derecha. La mayoría de países tienen valores entre 5 y 15, con pocos casos más altos.

pop15: Parece relativamente simétrica, quizá con dos agrupaciones ligeras ya que la mayoría de datos están en los extremos de la gráfica.

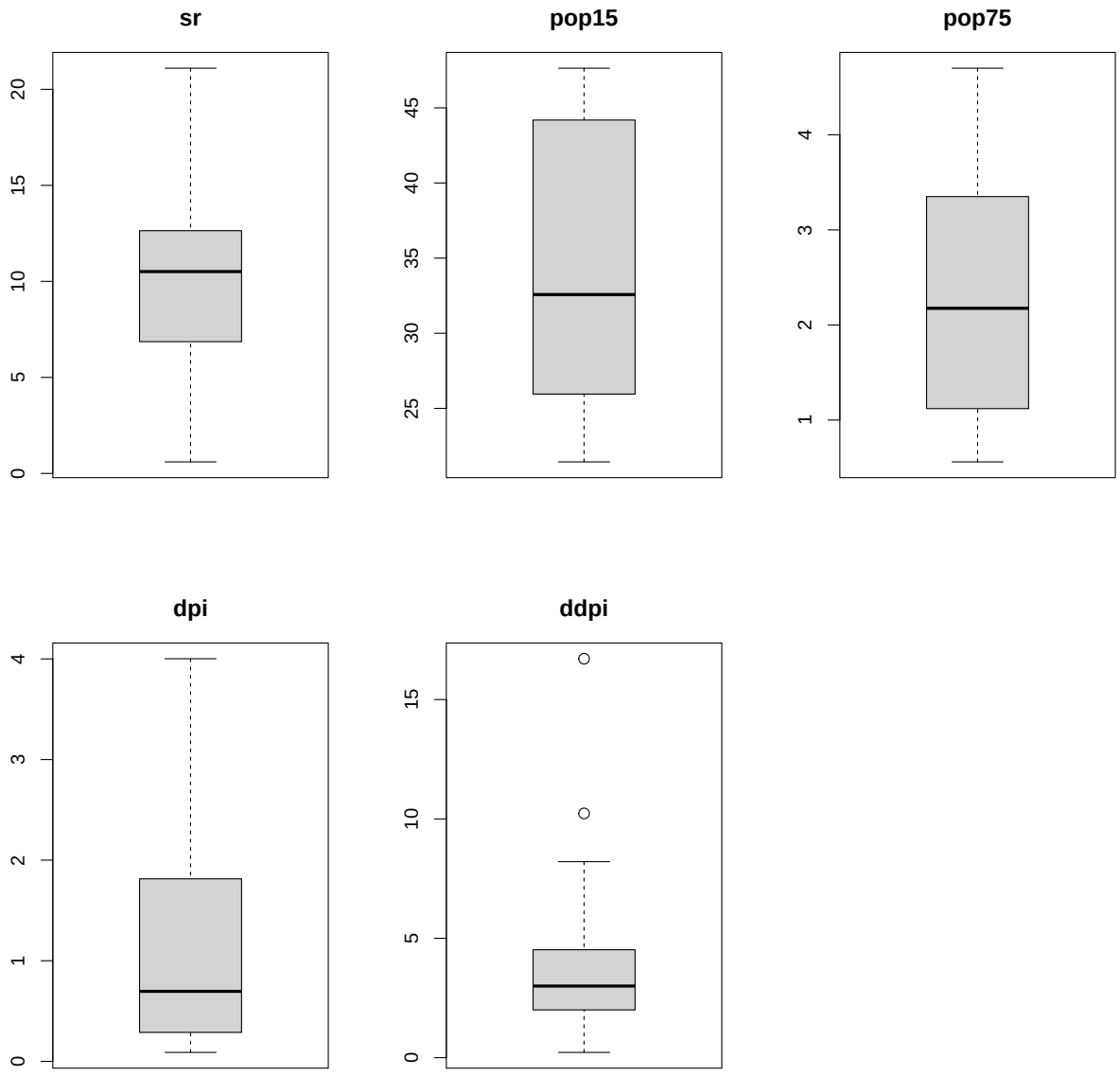
pop75: Distribución más dispersa. También con ligera cola derecha.

dpi: Muy sesgada a la derecha. La mayoría de valores son bajos, con algunos países muy por encima del resto (outliers).

ddpi: También sesgada a la derecha, aunque menos extrema que dpi. La mayor parte de valores están entre 0 y 5, con algunos casos altos en torno a más de 15 (outliers).

Boxplots

```r
par(mfrow = c(2,3),
    mar = c(4,4,3,1),
    cex = 1.2)
for(i in 1:5) boxplot(sv[,i],main=names(sv)[i])
```



sr: Distribución moderadamente simétrica, sin valores atípicos extremos. La mediana está cerca del centro del rango intercuartílico (IQR), con variabilidad moderada.

pop15: Mayor dispersión. La mediana está algo más cerca del cuartil inferior, indicando ligera asimetría hacia arriba. No se observan outliers.

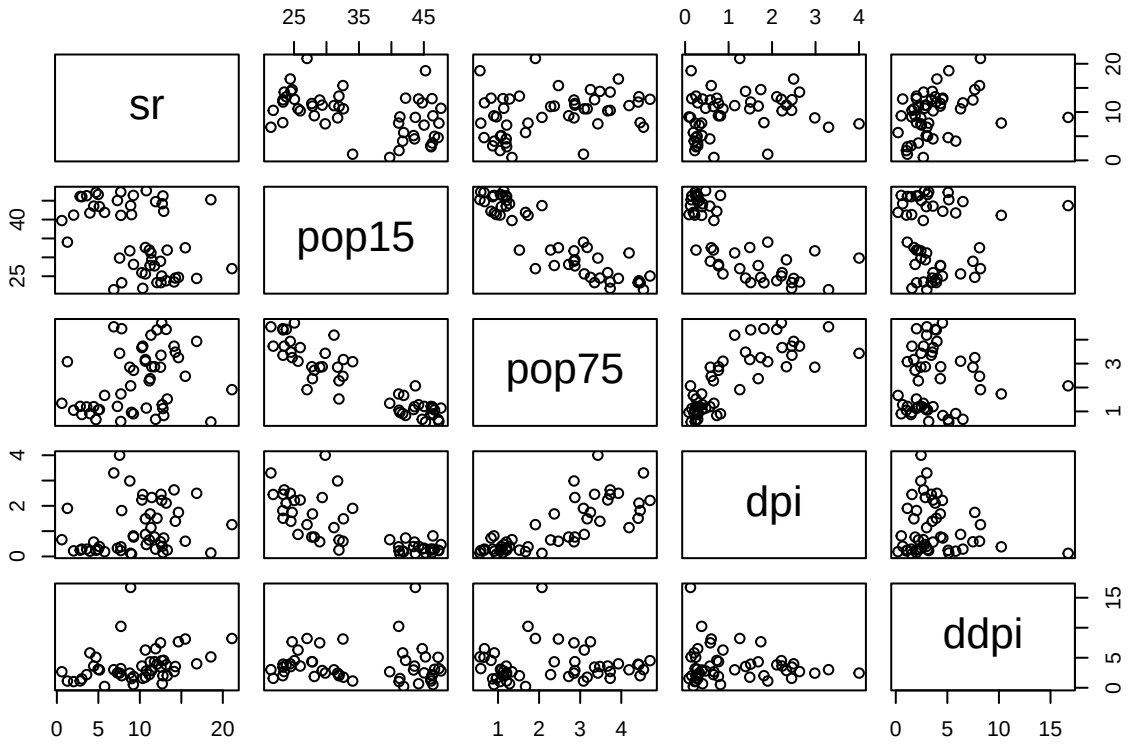
pop75: IQR relativamente estrecho y sin outliers. Los valores están bastante concentrados, con ligera asimetría hacia arriba (la cola superior más larga).

dpi: Fuertemente asimétrico. Gran distancia entre la mediana y el límite superior, lo que indica países con valores muy altos. No se muestran outliers explícitos, pero claramente hay una cola larga para valores más

altos.

ddpi: Variabilidad baja en general, pero con algunos outliers claros por encima de 10–15. La mediana está baja en el IQR, lo que indica sesgo a la derecha.

```
pairs(sv)
```



Destacar las relaciones lineales inversas entre pop15 y pop75. Tiene cierto sentido ya que cuanto mayor es la tasa de población joven suele haber menos población mayor. Esta relación se ve claramente entre países desarrollados y no desarrollados.

```
#RELACIÓN LINEAL DE LAS VARIABLES
cov(sv)
```

Covarianza y correlación

```
##                sr      pop15        pop75        dpi        ddpi
## sr     20.0740459 -18.678638   1.83049898  0.9782825  3.91901061
## pop15 -18.6786384  83.754110 -10.73166612 -6.8572360 -1.25610710
## pop75   1.8304990 -10.731666   1.66609082  1.0065607  0.09379918
## dpi     0.9782825  -6.857236   1.00656075  0.9818212 -0.36821351
## ddpi    3.9190106  -1.256107   0.09379918 -0.3682135  8.23615739
```

```r
cor(sv)
```

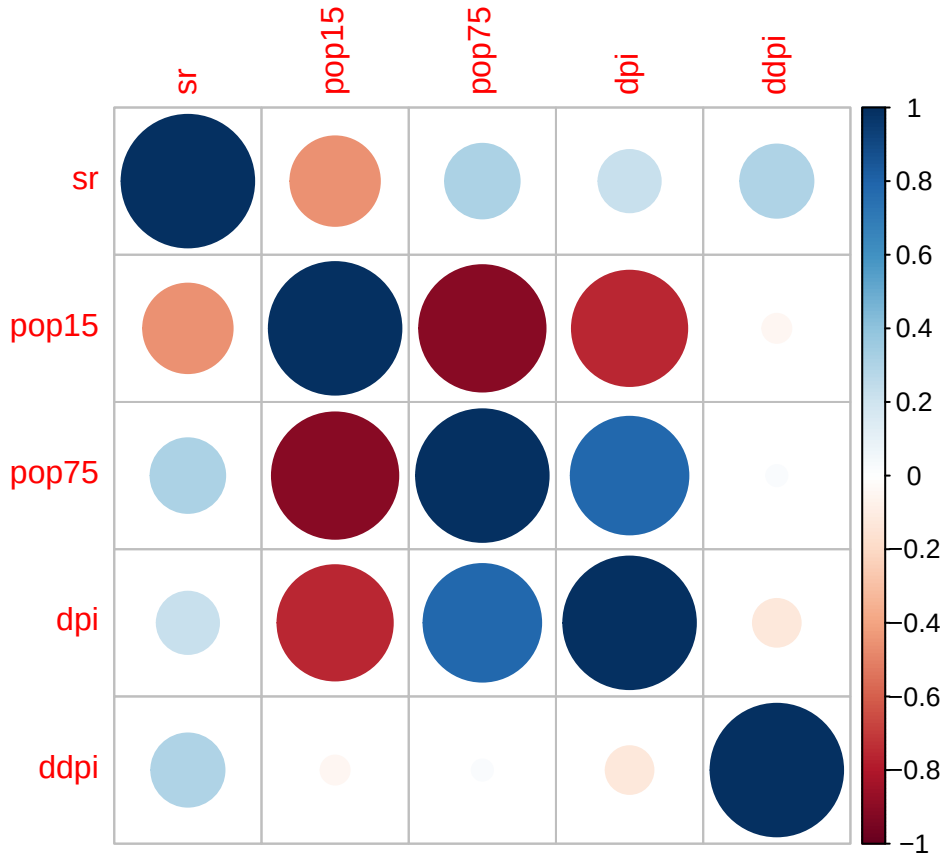
```
##               sr       pop15       pop75        dpi        ddpi
## sr     1.0000000 -0.45553809  0.31652112  0.2203589  0.30478716
## pop15 -0.4555381  1.00000000 -0.90847871 -0.7561881 -0.04782569
## pop75  0.3165211 -0.90847871  1.00000000  0.7869995  0.02532138
## dpi    0.2203589 -0.75618810  0.78699951  1.0000000 -0.12948552
## ddpi   0.3047872 -0.04782569  0.02532138 -0.1294855  1.00000000
```

La variable sr presenta una correlación negativa moderada con pop15, lo que indica que una mayor proporción de población joven se asocia con menores tasas de ahorro. También muestra correlaciones positivas moderadas con ddpi y pop75, sugiriendo que tanto el crecimiento de la renta disponible como un mayor envejecimiento de la población tienden a relacionarse con mayores niveles de ahorro. La relación con dpi es positiva pero débil.

Entre las variables explicativas destaca una fuerte correlación negativa entre pop15 y pop75, lo que refleja que los países con muchos jóvenes suelen tener pocos mayores y viceversa. Además, pop15 y dpi presentan correlación negativa elevada, mientras que pop75 y dpi muestran correlación positiva alta, lo que indica posible multicolinealidad entre las variables demográficas y el nivel de renta. Por su parte, ddpi tiene correlaciones muy bajas con el resto, por lo que actúa como predictor prácticamente independiente.

- Gráfica de correlación

```r
#Vemos la correlación gráficamente
par(mfrow=c(1,1))
library('corrplot')
corrplot(cor(sv),method="circle")
```



Con esta gráfica se refuerza lo ya dicho anteriormente:

- pop15 y pop75 están fuertemente correlacionados negativamente, lo que tiene sentido porque representan proporciones de población joven y mayor (cuando una sube, la otra baja).
- dpi y pop75 muestran una correlación positiva relativamente fuerte.
- ddpi parece tener correlaciones muy débiles con casi todas las variables.

FASE III

Regresión múltiple del modelo Describir el modelo ajustado y sus residuos, contrastar la significatividad individual de los parámetros y la calidad del modelo, verificar las hipótesis del modelo, análisis de datos influyentes y atípicos

Primero se obtiene el objeto de regresión múltiple y se analiza.

```
gc <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = sv)
summary(gc)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = sv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.5661     7.3545   3.884 0.000334 ***
## pop15        -0.4612     0.1446  -3.189 0.002603 **
## pop75        -1.6915     1.0836  -1.561 0.125530
## dpi          -0.3369     0.9311  -0.362 0.719173
## ddpi          0.4097     0.1962   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

La tasa de ahorro sr se ve explicada principalmente por dos variables:

-pop15 (negativa y bastante significativa)
-ddpi (positiva y significativa)

Las variables pop75 y dpi no resultan significativas dentro del modelo (p-value superior a 0.05), probablemente a causa de multicolinealidad entre las variables demográficas y el nivel de renta.

El modelo es globalmente significativo ya que el F-statistic sugiere que al menos una de sus variables predictivas tiene más capacidad explicativa que un modelo sin predictores ($p < 0.001$), pero su capacidad explicativa es limitada ($R\hat{}2 = 0.34$). Se deberá depurar el modelo eliminando variables no significativas.

```
anova(gc)
```

Análisis de la tabla ANOVA

```
## Analysis of Variance Table
##
## Response: sr
##           Df Sum Sq Mean Sq F value    Pr(>F)
## pop15      1 204.12 204.118 14.1157 0.0004922 ***
## pop75      1  53.34  53.343  3.6889 0.0611255 .
## dpi        1  12.40  12.401  0.8576 0.3593551
## ddpi       1  63.05  63.054  4.3605 0.0424711 *
## Residuals 45 650.71  14.460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La tabla ANOVA proporciona información sobre la explicabilidad de los resultados de sr en función de cada variable predictora.
Según los resultados de la tabla ANOVA, pop15 y ddpi son las variables con más capacidad predictora, pop75 es ligeramente no significativo y dpi es completamente insignificativo.

Estudio de la multicolinealidad Se recuerda que la multicolinealidad puede inflar las varianzas y los pesos de las variables, haciendo que un pequeño cambio en el dataset ocasione un gran cambio en las rectas de regresión. Para estudiarla se calcula la matriz de autovalores.

```r
X <- model.matrix(gc)[, 2:5]    # quitamos el intercepto
R <- cor(X)

# Autovalores de la matriz de correlaciones
eig <- eigen(R)

# Índice de condición para cada autovalor
k <- sqrt(max(eig$values) / eig$values)
k
```

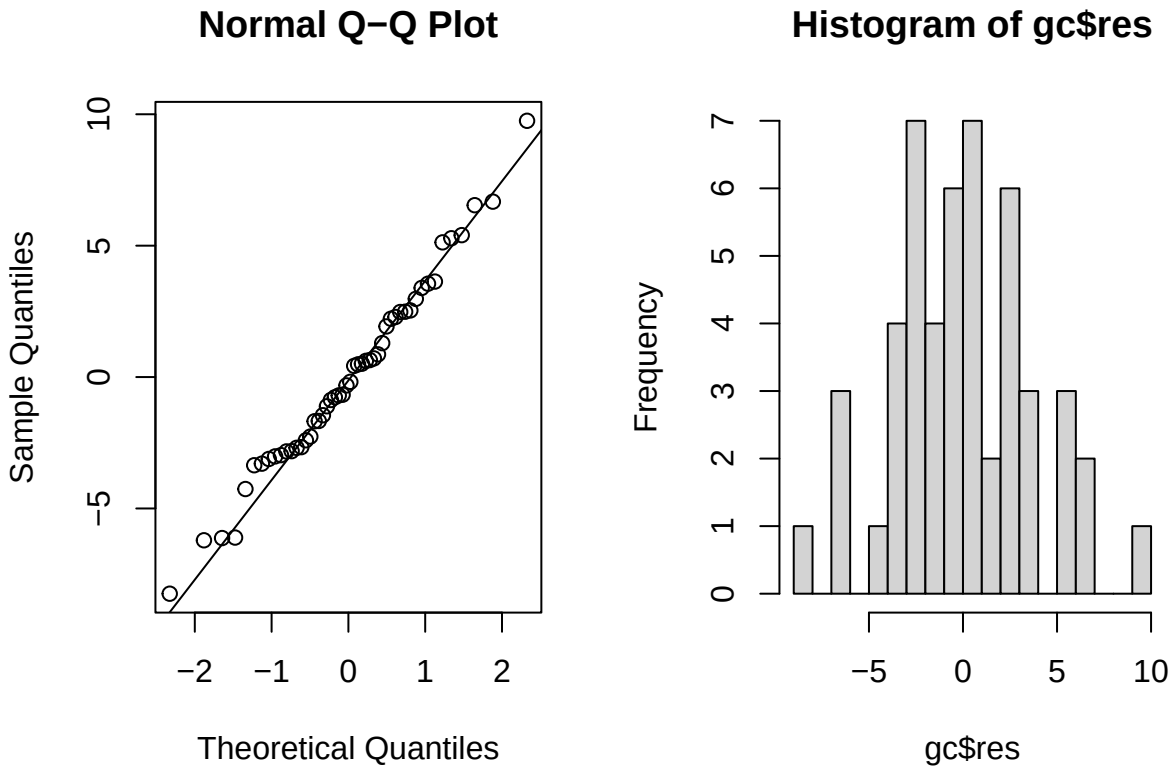
```
## [1] 1.000000 1.604196 3.254441 5.423014
```

Ninguno es supeior a 10 por lo que no parece haber clara colinealidad..

Comprobación de normalidad, independencia, homocedasticidad, linealidad etc.

- *Normalidad de residuos*

```r
#Normalidad: QQ-plot de los residuos e histograma
par(mfrow=c(1,2))
qqnorm(gc$res)
qqline(gc$res)
hist(gc$res,13)
```



No es una normal perfecta pero sí que sigue bastante el patrón de una normal, por lo que no se descarta normalidad a priori. Puede ser que con más datos la forma fuera algo más perfecta.

Contraste de `Saphiro-Wilk` para *normalidad de residuos*. Se usa este porque solo se dispone de 50 filas de datos.

```
shapiro.test(gc$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gc$res
## W = 0.98698, p-value = 0.8524
```

No podemos rechazar la hipótesis nula (los residuos tienen una distribución normal)

- *Independencia con Durbin-Watson*

```
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dwtest(gc,alternative ="two.sided",iterations = 1000)
```

```
##
##  Durbin-Watson test
##
## data:  gc
## DW = 1.9341, p-value = 0.7794
## alternative hypothesis: true autocorrelation is not 0
```

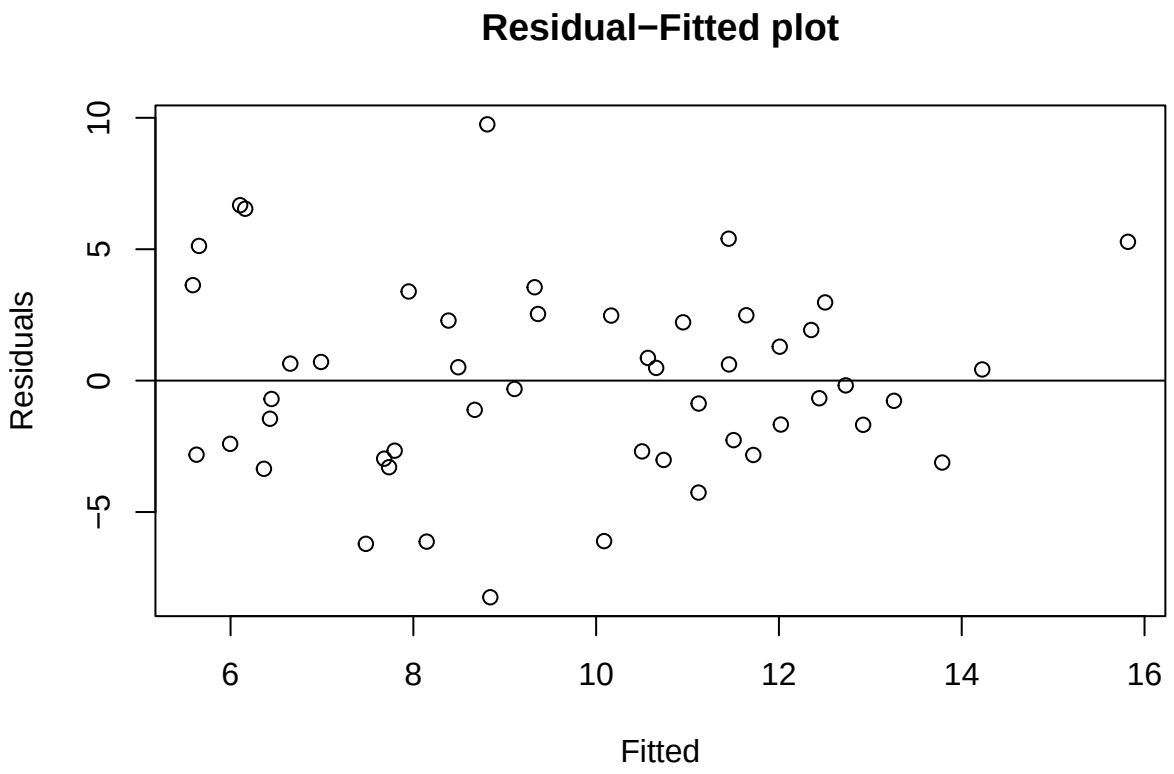
El estadístico es muy cercano a 2, lo que indica ausencia de autocorrelación en residuos.

No podemos rechazar la hipótesis de que los residuos no son independientes (p-value alto)

- *Homocedasticidad*

Gráfica de residuos

```
#Muestro los residuos del modelo gráficamente
par(mfrow=c(1,1))
plot(gc$fit,gc$res,xlab="Fitted",ylab="Residuals", main="Residual-Fitted plot")
abline(h=0)
```



En la gráfica ya se aprecia una varianza constante de los residuos. Destaca algún residuo por su magnitud. De igual forma, se comprueba con contraste `Breusch-Pagan`

```
bptest(gc)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  gc
## BP = 4.9852, df = 4, p-value = 0.2888
```

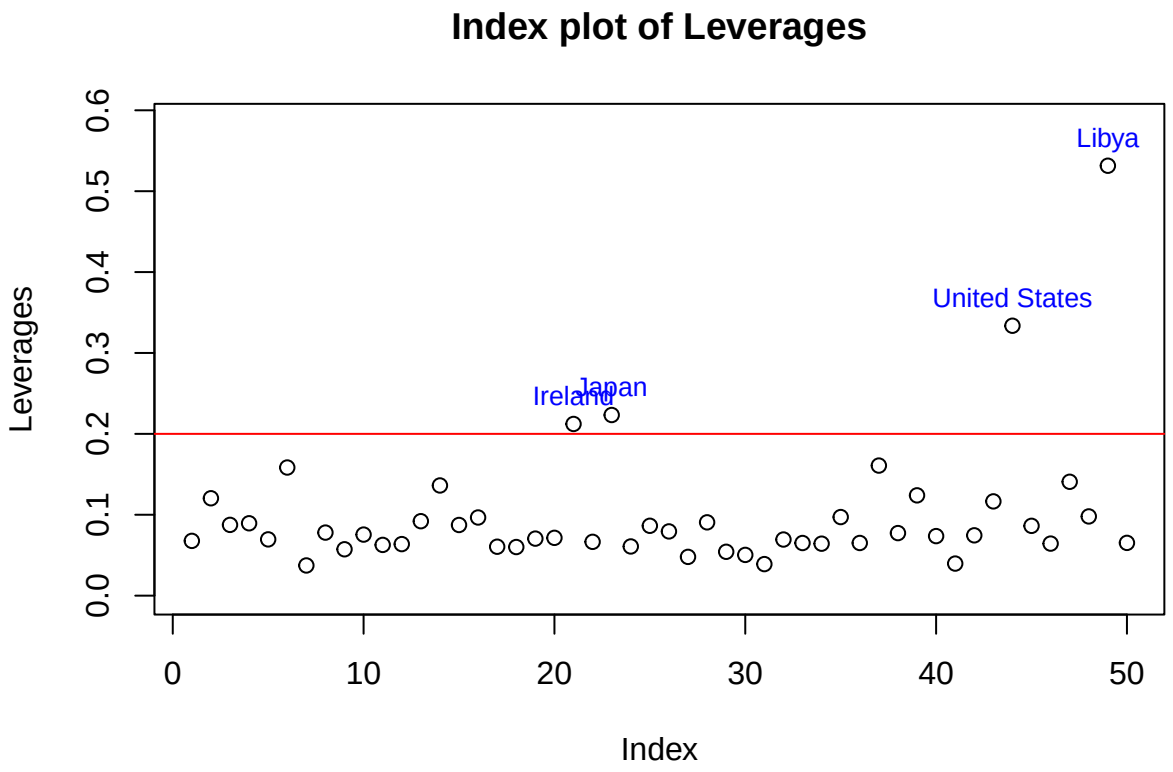
El p-value es mucho mayor a 0.05 por lo que no se puede rechazar la hipótesis nula, es decir, que la varianza de los residuos sea constante.

Análisis de puntos palanca e influyentes

Estudio de puntos palanca

```
x <- model.matrix(gc)
leverageC <- hat(x)

# Definir la línea de referencia
h <- 2 * sum(leverageC) / 50

# Ajustar ylim para que haya espacio arriba
```

```r
ylim_max <- max(leverageC) * 1.1  # 10% más arriba para no cortar un punto
plot(leverageC, ylab = "Leverages", main = "Index plot of Leverages", ylim = c(0, ylim_max))

abline(h = h, col = "red")

outliers <- which(leverageC > h)
text(x = outliers, y = leverageC[outliers], labels = rownames(savings)[outliers],
     pos = 3, cex = 0.8, col = "blue")
```



Se aprecia que Irlanda, Japón, Libya y EEUU superan el umbral. Esto quiere decir que estos puntos son potencialmente influyentes en la recta de regresión.

```r
#Muestro los efectos palanca de estos puntos
names(leverageC) <- names
leverageC[leverageC > 2*sum(leverageC)/50]
```

```
##       Ireland         Japan United States         Libya
##     0.2122363     0.2233099     0.3336880     0.5314568
```

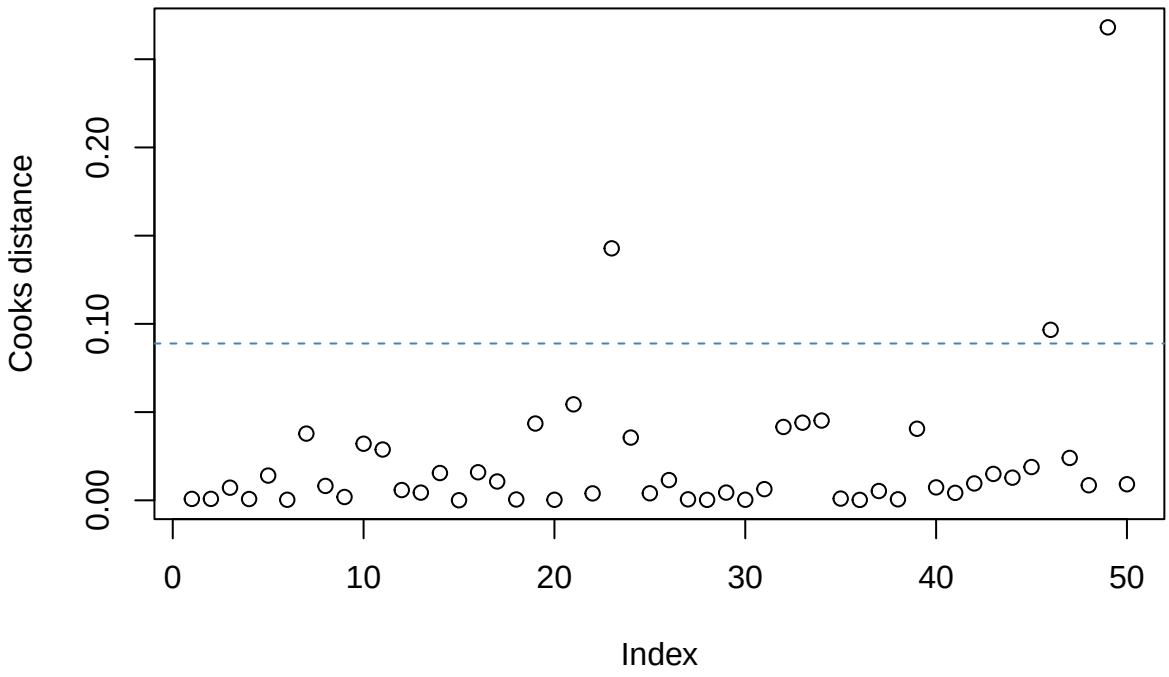
Se procede a estudiar la *distancia de Cook*.

La distancia de Cook es otra medida que indica cuánto influye cada observación individual en los resultados de un modelo de regresión.

- Combina residuos y apalancamiento.

- Un valor alto significa que esa observación tiene un impacto grande en los coeficientes del modelo.
- Se usa para detectar outliers influyentes que podrían distorsionar la regresión.

```
#Identificación de puntos influyentes (estadístico de Cook)
cookC <- cooks.distance(gc)
plot(cookC,ylab="Cooks distance")
abline (h = 4 / (50-4-1), lty = 2, col = "steelblue")
```




```
# Umbral
threshold <- 4 / (50-4-1)

# Índices de observaciones con Cook alto
high_cook <- which(cookC > threshold)
high_cook
```

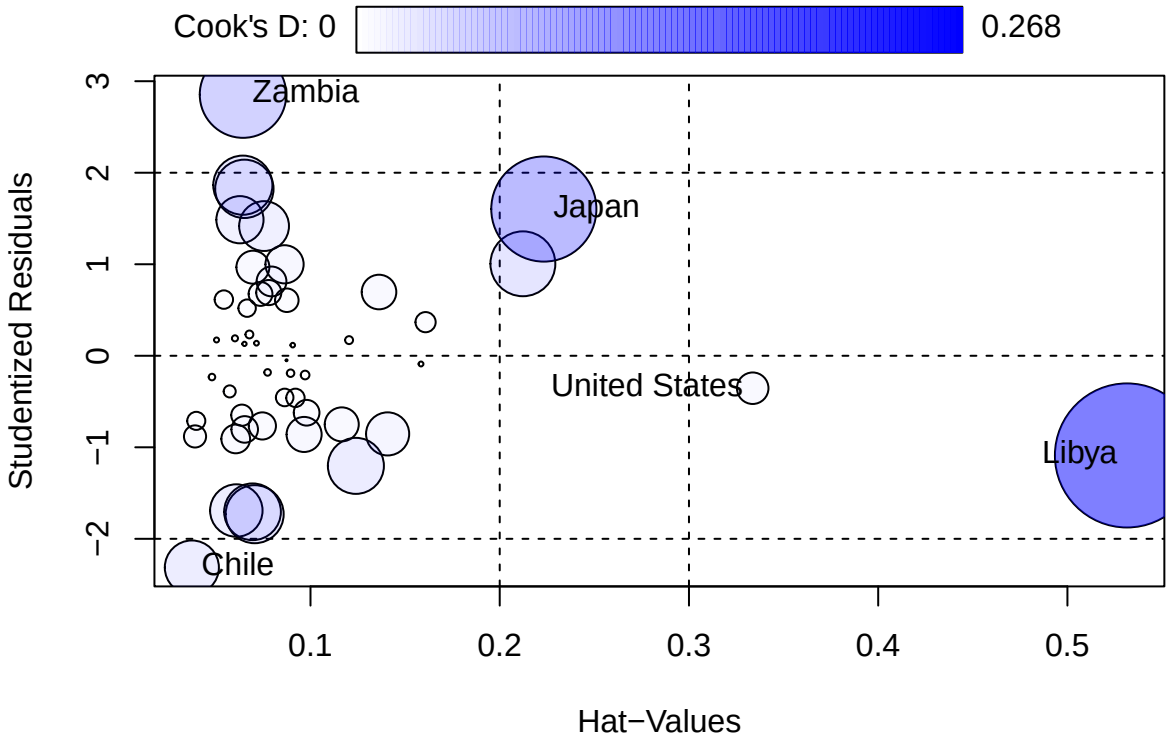
```
##  Japan Zambia  Libya
##     23     46     49
```

En este caso, Japan, Zambia y Lybia parecen ser los puntos que superan el threshold, es decir, aquellos potencialmente influyentes.

InfluencePlot para analizar efecto palanca.

```r
#Otra forma de analizar el efecto palanca, los puntos influyentes y los valores atípicos es usando la f
library(car)
influencePlot(gc, id.method="identify")
```

```
## Warning in plot.window(...): "id.method" es un parámetro gráfico inválido

## Warning in plot.xy(xy, type, ...): "id.method" es un parámetro gráfico inválido

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" es un
## parámetro gráfico inválido
## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" es un
## parámetro gráfico inválido

## Warning in box(...): "id.method" es un parámetro gráfico inválido

## Warning in title(...): "id.method" es un parámetro gráfico inválido

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" es un
## parámetro gráfico inválido
```



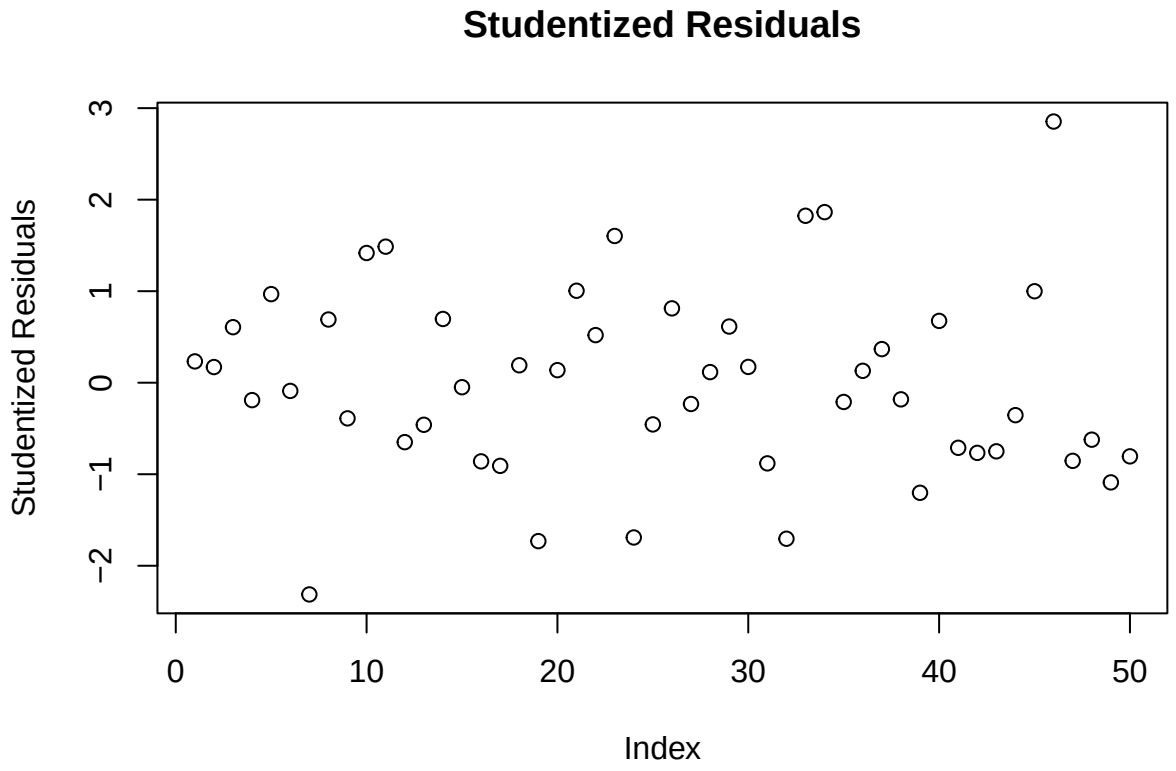

```
##                StudRes        Hat      CookD
## Chile       -2.3134295 0.03729796 0.03781324
## Japan        1.6032158 0.22330989 0.14281625
```

```
## United States -0.3546151 0.33368800 0.01284481
## Zambia         2.8535583 0.06433163 0.09663275
## Libya         -1.0893033 0.53145676 0.26807042
```

Se aprecia que Lybia, Japan y Zambia destacan respecto al resto. Este plot es útil para estudiar qué es lo que hace destacar a los puntos, si su efecto palanca (hat-values) o el tamaño del residuo estudentizado.

Estudio de residuos estudentizados

```
studC<- rstudent(gc)
plot(studC,ylab="Studentized Residuals",main="Studentized Residuals")
```



Ordenamos los códigos postales por su residuo estudentizado.

```
sort(abs(studC))
```

```
##        Germany         Canada    Netherlands   South Africa          India
##     0.04918692     0.08983197     0.11605663     0.12996586     0.13729730
##        Austria      Nicaragua          Spain        Bolivia       Honduras
##     0.17095506     0.17254242     0.18175853     0.19037831     0.19051919
##       Portugal         Norway      Australia  United States South Rhodesia
##     0.21040432     0.23247367     0.23271611     0.35461507     0.36714512
##       Colombia     Luxembourg        Finland          Italy        Belgium
##     0.38946778     0.45560591     0.45986445     0.52015744     0.60655220
##    New Zealand        Uruguay        Ecuador    Switzerland          China
##     0.61373189     0.62253411     0.64957871     0.67532922     0.69048169
```

```
##         France          Turkey United Kingdom         Tunisia        Malaysia
##     0.69640933     0.71138840     0.74959873     0.76677907     0.80489153
##          Malta        Jamaica         Greece         Panama      Guatamala
##     0.81227407     0.85376418     0.85967533     0.88147653     0.90854545
##         Brazil      Venezuela        Ireland          Libya         Sweden
##     0.96790816     0.99932569     1.00485886     1.08930326     1.20293404
##     Costa Rica        Denmark          Japan          Korea       Paraguay
##     1.41731062     1.48644473     1.60321582     1.69103214     1.70488128
##        Iceland           Peru    Philippines          Chile         Zambia
##     1.73119989     1.82391409     1.86382587     2.31342946     2.85355834
```

```r
# t_(n-k-1,1-alfa/2) con nivel de significación alfa=0.05
qt(0.975,45)
```

```
## [1] 2.014103
```

Como se puede ver, los residuos estudentizados más altos no tienen por qué proceder de los puntos palanca. Según el valor crítico de la distribución t de student, a partir de 2.014 se consideran outliers. Chile y Zambia son dos posibles candidatos.

En caso de quedarnos con este modelo se decidiría eliminar aquellos puntos con distancia de Cook mayor al threshold (Japon, Zambia y Lybia).

FASE IV

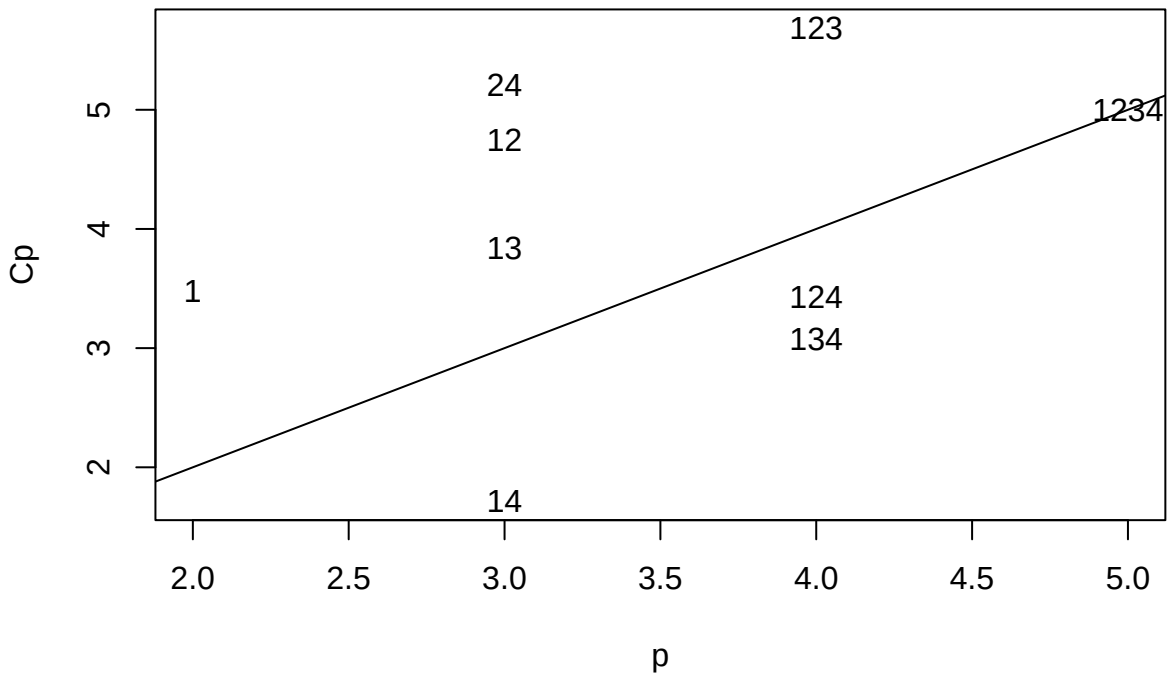
Identificar el mejor modelo de regresión lineal y analizarlo. Se procede a implementar distintos métodos para detectar qué modelo es el mejor. Se trata de buscar un modelo lo más simple posible que tenga una capacidad predictora competente.

Estadístico `Cp de Mallow`

```r
cook_filter <- cooks.distance(gc) < threshold

y <- savings$sr[cook_filter]
x <- savings[cook_filter, !names(savings) %in% "sr"]

# Selección de variables
library(leaps)
```

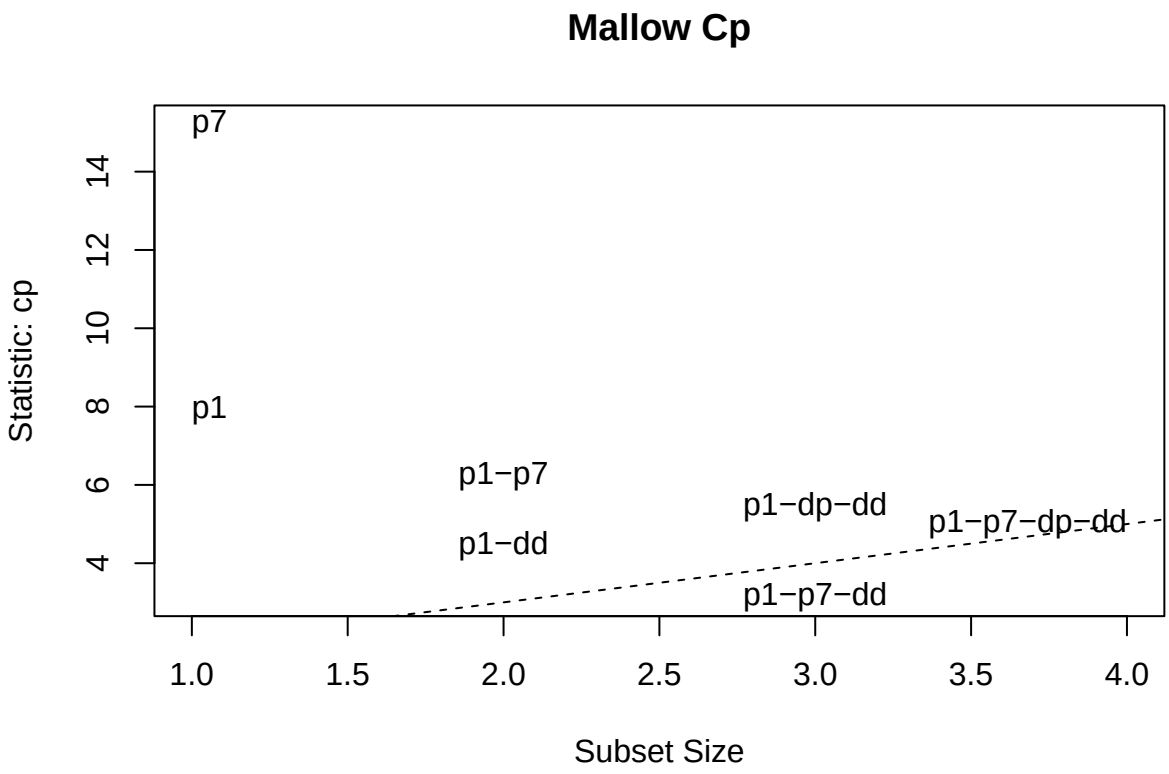
```
## Warning: package 'leaps' was built under R version 4.5.2
```

```r
subset_models <- leaps(x, y)

# Graficar Cp de Mallows
Cpplot(subset_models)
```



Según el `Cp de Mallow`, los dos modelos más interesantes serían el 124 y el 14. Si bien es cierto que el 124 se acerca más a la linea por debajo, el 14 es más simple y a su vez está por debajo de la línea, lo que indica que se ajusta bien a los datos.

```
modSS <- regsubsets(sr ~ ., data = sv, nbest = 2, intercept = TRUE)
mallowCp <- subsets(modSS, statistic="cp", legend = FALSE, min.size = 1, main = "Mallow Cp")
abline(a = 1, b = 1, lty = 2)
```



Esta es otra forma de representar la gráfica del `Cp de Mallow`. Se reitera lo dicho anteriormente, los modelos candidatos serían el de variables 14 y 124.

`R cuadrado ajustado`:
Premia la capacidad explicativa y penaliza el exceso de predictores que solo aumentan el ruido.

```
adjr <- leaps(x,y,method="adjr2")
maxadjr(adjr,4)
```

```
##     1,4   1,3,4   1,2,4 1,2,3,4
##   0.309   0.304   0.298   0.288
```

Aquí se ve como el mejor R^2 ajustado lo tiene el modelo más simple de predictores 1 y 4

Selección con `AIC` usando *stepwise*

```
back <- lm(sr ~ ., sv)
sm <- step(back, direction = "both")
```

```
## Start:  AIC=138.3
## sr ~ pop15 + pop75 + dpi + ddpi
##
##         Df Sum of Sq    RSS    AIC
## - dpi    1     1.893 652.61 136.45
## <none>               650.71 138.30
## - pop75  1    35.236 685.95 138.94
```

```
## - ddpi   1    63.054 713.77 140.93
## - pop15  1   147.012 797.72 146.49
##
## Step:  AIC=136.45
## sr ~ pop15 + pop75 + ddpi
##
##         Df Sum of Sq    RSS    AIC
## <none>               652.61 136.45
## - pop75  1    47.946 700.55 137.99
## + dpi    1     1.893 650.71 138.30
## - ddpi   1    73.562 726.17 139.79
## - pop15  1   145.789 798.40 144.53
```

El algoritmo ha eliminado la variable dpi porque disminuye el AIC y elige como mejor modelo el compuesto por las variables pop75, pop15 y ddpi. Destacar también que en caso de quitar la variable pop75 la penalización es relativamente baja (vease el AIC en la tabla de abajo y fila -pop75) y la ventaja es la simplificación del modelo.

Para la *decisión final* se tiene en cuenta lo siguiente:
- Los modelos candidatos eran los de variables pop15 y ddpi y las mismas junto con pop75. Debido a que ambos modelos son prácticamente iguales en cuanto a capacidad de predicción y además obtienen puntuaciones parecidas en AIC y R^2 ajustado, se decide escoger el modelo más simple compuesto por dos variables predictoras: *pop15 y ddpi*

```
mfin <- lm(sr ~ pop15 + ddpi, sv) # modelo final
summary(mfin)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + ddpi, data = sv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5831 -2.8632  0.0453  2.2273 10.4753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.59958    2.33439   6.682 2.48e-08 ***
## pop15       -0.21638    0.06033  -3.586 0.000796 ***
## ddpi         0.44283    0.19240   2.302 0.025837 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.861 on 47 degrees of freedom
## Multiple R-squared:  0.2878, Adjusted R-squared:  0.2575
## F-statistic: 9.496 on 2 and 47 DF,  p-value: 0.0003438
```

Parece que el modelo actual es prácticamente igual de bueno que el anterior pese a haber reducido el número de predictores. Dichos predictores tienen ahora p-value inferior a 0.05 por lo que se consideran influyentes en el resultado. El F-statistic sigue siendo inferior a 0.001.

Análisis de la tabla ANOVA

```
anova(mfin)
```

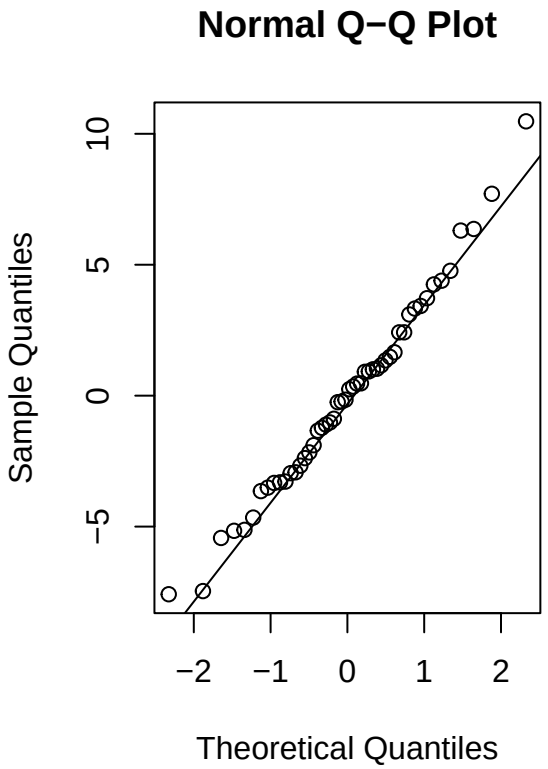
```
## Analysis of Variance Table
##
## Response: sr
##           Df Sum Sq Mean Sq F value    Pr(>F)
## pop15      1 204.12 204.118 13.6942 0.0005633 ***
## ddpi       1  78.96  78.959  5.2973 0.0258374 *
## Residuals 47 700.55  14.905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

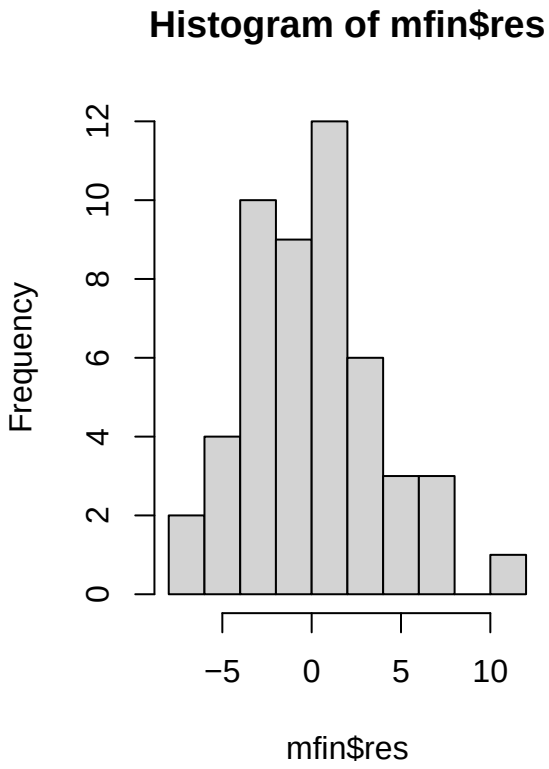
Ambos p-values son menores que 0.05, lo cual es positivo. Además, la variable predictora más importante es pop15 aunque el ddpi sí se considera algo importante y por tanto, debe permanecer.

Supuestos del modelo:

- *Normalidad*

```
#Normalidad: QQ-plot de los residuos e histograma
par(mfrow=c(1,2))
qqnorm(mfin$res)
qqline(mfin$res)
hist(mfin$res,10)
```





El gráfico QQplot ya nos indica una clara normalidad en los residuos, ya que los puntos siguen relativamente bien la linea salvo algún punto muy concreto. El histograma también tiene cierto parecido con el de una normal salvo por unos outliers muy marcados con valores superiores a 10. Se realiza el contraste the `Saphiro-Wilk`para cerciorarse de la normalidad de los residuos.

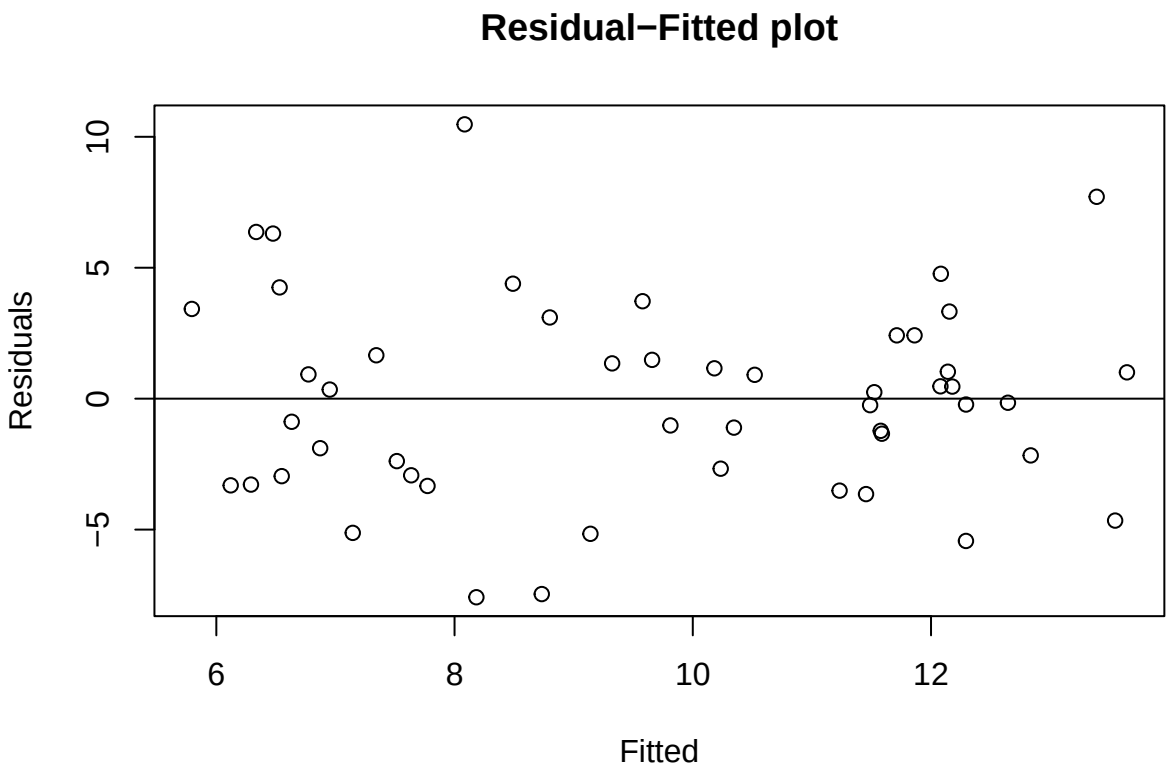
```
shapiro.test(mfin$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mfin$res
## W = 0.98666, p-value = 0.8398
```

No podemos rechazar que los datos provengan de una normal. (p-value muy alto)

- *Homocedasticidad*
  Análisis de Residuos

```
plot(mfin$fit, mfin$res,xlab="Fitted",ylab="Residuals", main="Residual-Fitted plot")
abline(h=0)
```



Parece que sí hay homocedasticidad, ya que se aprecia una varianza constante en los datos. Igualmente se comprueba con el contraste de `Breusch-Pagan`

```
bptest(mfin)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mfin
## BP = 3.5528, df = 2, p-value = 0.1692
```

El p-value es superior a 0.05 por lo que no podemos rechazar la homocedasticidad de los residuos

- *Independencia*
  Estadístico de `Durbin-Watson`

```
#Independencia (estadístico de Durbin-Watson)
dwtest(mfin,alternative ="two.sided",iterations = 1000)
```

```
##
##  Durbin-Watson test
##
## data:  mfin
## DW = 2.0324, p-value = 0.9381
## alternative hypothesis: true autocorrelation is not 0
```

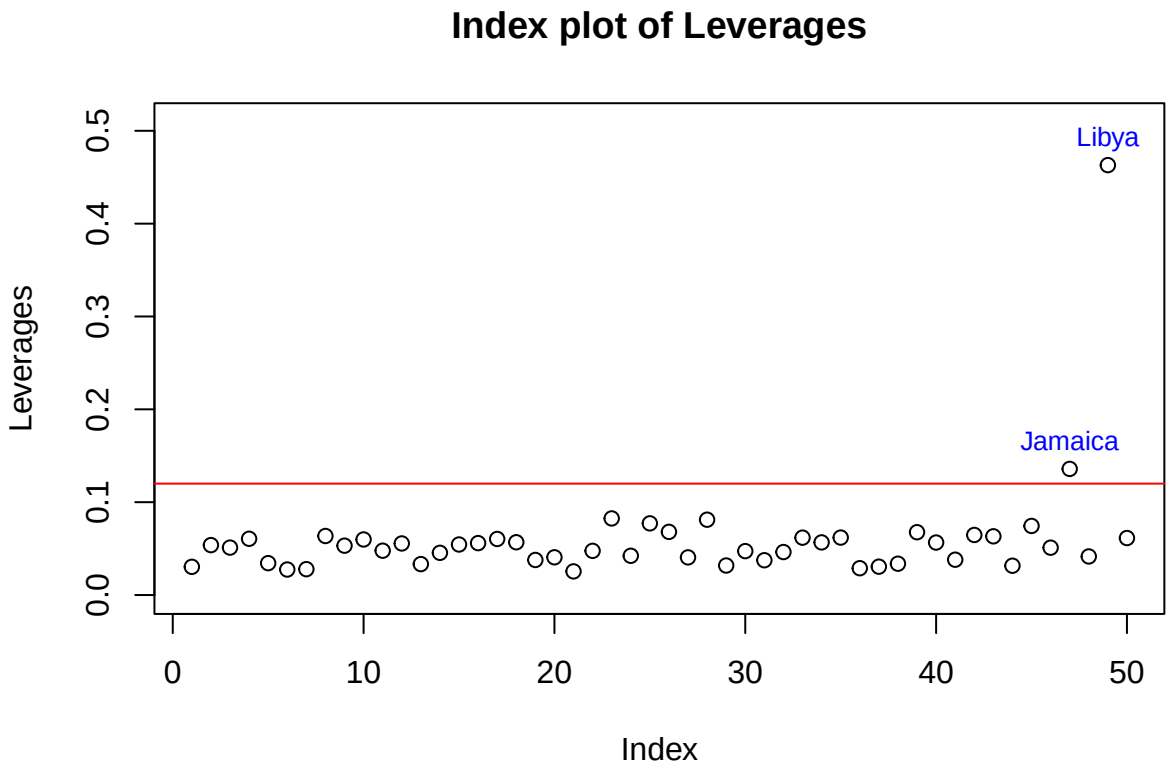
No podemos rechazar la hipótesis de que los residuos no son independientes. Es por ello que no se considera correlación entre residuos.

El modelo cumple los supuestos y, por tanto, se puede llevar a cabo el estudio de puntos palanca e influyentes para mejorarlo.

Análisis de puntos palanca e influyentes

Puntos palanca

```
x <- model.matrix(mfin)
leverageC <- hat(x)

# Definir la línea de referencia
h <- 2 * sum(leverageC) / 50

# Ajustar ylim para que haya espacio arriba
ylim_max <- max(leverageC) * 1.1  # 10% más arriba para no cortar un punto
plot(leverageC, ylab = "Leverages", main = "Index plot of Leverages", ylim = c(0, ylim_max))

abline(h = h, col = "red")

outliers <- which(leverageC > h)
text(x = outliers, y = leverageC[outliers], labels = rownames(savings)[outliers],
     pos = 3, cex = 0.8, col = "blue")
```



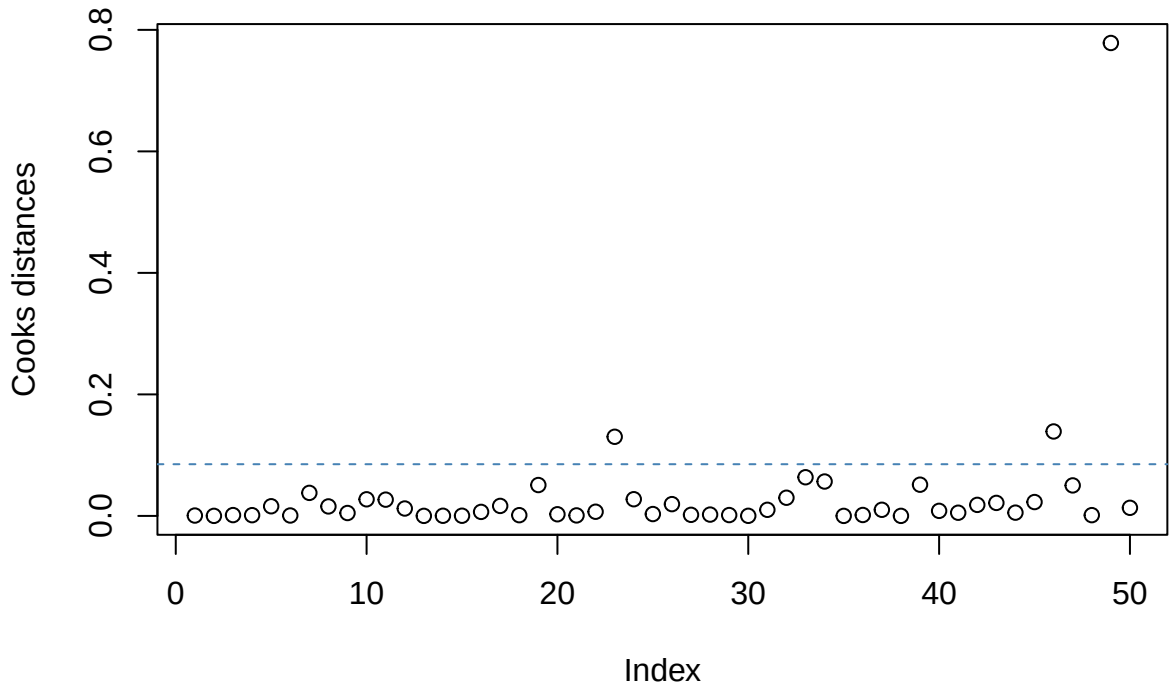
Se aprecian como puntos influyentes Lybia y Jamaica. Dejamos estos dos puntos como candidatos a ser eliminados. Cabe destacar que Lybia tiene un efecto palanca muy superior a Jamaica.

```
#Muestro los efectos palanca de estos puntos
names(leverageC) <- names
leverageC[leverageC > 2*sum(leverageC)/50]
```

```
##   Jamaica     Libya
## 0.1358301 0.4630158
```

Distancia de Cook

```
#Identificación de puntos influyentes (estadístico de Cook)
cookC <- cooks.distance(mfin)
plot(cookC,ylab="Cooks distances")
abline (h = 4 / (50-2-1), lty = 2, col = "steelblue")
```




```
# Umbral
threshold <- 4 / (50-2-1)

# Índices de observaciones con Cook alto
high_cook <- which(cookC > threshold)
high_cook
```

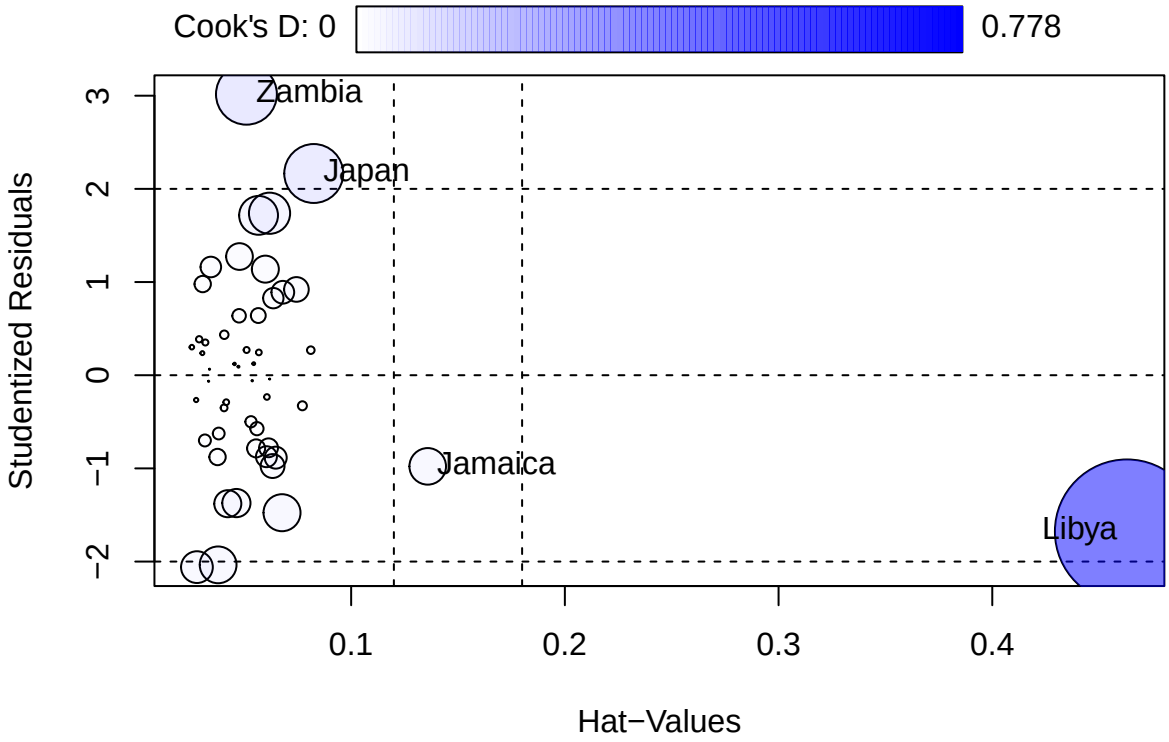
```
##  Japan Zambia  Libya
##     23     46     49
```

Japón, Zambia y Lybia son los 3 países con mayor distancia de Cook.

Puntos de influencia con `InfluencePlot`

```
influencePlot(mfin, id.method="identify")
```

```
## Warning in plot.window(...): "id.method" es un parámetro gráfico inválido

## Warning in plot.xy(xy, type, ...): "id.method" es un parámetro gráfico inválido

## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" es un
## parámetro gráfico inválido
## Warning in axis(side = side, at = at, labels = labels, ...): "id.method" es un
## parámetro gráfico inválido

## Warning in box(...): "id.method" es un parámetro gráfico inválido

## Warning in title(...): "id.method" es un parámetro gráfico inválido

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.method" es un
## parámetro gráfico inválido
```




```
##            StudRes        Hat      CookD
## Japan    2.1648444 0.08249690 0.13024687
```

```
## Zambia   3.0155346 0.05100486 0.13898018
## Jamaica -0.9781968 0.13583006 0.05017955
## Libya   -1.6770997 0.46301580 0.77838805
```

Zambia, Japan y Lybia son los puntos a destacar en el gráfico. Efectivamente, Lybia es el que mayor hat-value tiene, y Zambia y Japan son puntos que destacan porque sus residuos estudentizados son outliers. Vease mejor en el análisis de residuos estudentizados.

Análisis de Residuos estudentizados

```
#Comprobemos que efectivamente son datos atípicos a partir del análisis de los residuos estudentizados
studC<- rstudent(mfin)
sort(abs(studC))
```

```
##       Portugal        Austria          Spain        Finland      Nicaragua
##     0.04119097     0.05901078     0.06423212     0.06481754     0.09131325
##         France        Germany        Bolivia      Australia       Honduras
##     0.12098575     0.12408938     0.23358627     0.23699166     0.24467540
##         Canada    Netherlands        Belgium        Uruguay        Ireland
##     0.26581496     0.26886603     0.27078426     0.28960379     0.30104845
##     Luxembourg         Norway    New Zealand   South Africa          India
##     0.32790066     0.35040457     0.35127253     0.38590664     0.43437987
##       Colombia         Greece         Turkey          Italy    Switzerland
##     0.49951012     0.57353427     0.62583888     0.63778428     0.64060124
##  United States       Malaysia        Ecuador          China      Guatamala
##     0.70006990     0.77903623     0.78542283     0.82709092     0.87446788
##         Panama        Tunisia          Malta      Venezuela United Kingdom
##     0.87776204     0.88464187     0.89027851     0.92075864     0.97484220
##        Jamaica South Rhodesia     Costa Rica         Brazil        Denmark
##     0.97819684     0.97855441     1.13853682     1.16142021     1.27378525
##       Paraguay          Korea         Sweden          Libya    Philippines
##     1.37209315     1.37898330     1.47565073     1.67709968     1.71568572
##           Peru        Iceland          Chile          Japan         Zambia
##     1.73842322     2.03521378     2.05965764     2.16484438     3.01553465
```

```
qt(0.975,47)
```

```
## [1] 2.011741
```

Los residuos estudentizados más altos los obtienen Iceland, Chile, Japan y Zambia. Como analizamos con anterioridad, Japan y Zambia son los outiers mas destacables.

Conclusiones

Primeramente se vio como Lybia era el mayor leverage point aunque no era el residuo más grande. Luego, en el estudio de residuos estudentizados se aprecia como Zambia y Japón son los más grandes aunque no eran justamente los que más leverage tenían. La *distancia de Cook* tiene ambos factores en cuenta y es por ello que será la metrica en la que nos basemos para descartar puntos.

Los puntos a eliminar son Japan, Zambia y Lybia.

FASE V

Análisis del mejor modelo tras eliminar valores influyentes y atípicos. Eliminación de puntos con mayor distancia de Cook

```
# Filtro
sv_new <- sv[-high_cook, ]
# Se comprueba que se han eliminado dichos puntos
nrow(sv)
```

```
## [1] 50
```

```
nrow(sv_new)
```

```
## [1] 47
```

Estudio del mejor modelo sin dichos puntos

```
mfin_clean <- lm(sr ~ pop15 + ddpi, sv_new) # modelo final
summary(mfin_clean)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + ddpi, data = sv_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2720 -2.4965  0.0392  1.9016  6.7376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.23870    2.27244   6.706 3.09e-08 ***
## pop15       -0.21707    0.05541  -3.917 0.000308 ***
## ddpi         0.47179    0.23952   1.970 0.055180 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.39 on 44 degrees of freedom
## Multiple R-squared:  0.3393, Adjusted R-squared:  0.3092
## F-statistic:  11.3 on 2 and 44 DF,  p-value: 0.0001098
```

Se aprecia como ha aumentado en cierta medida el R^2 y el R^2 ajustado. Además, eliminando los puntos con mayor distancia de Cook, la variable ddpi ha perdido algo de relevancia ya que ahora su p-value es ligeramente superior a 0.05.

Ahora podemos presenciar un modelo simple (solo dos variables predictoras) pero muy competente con el inicial a la hora de predecir, ya que su R^2 ajustado ha mejorado. Es por ello que se considera este modelo como el mejor.